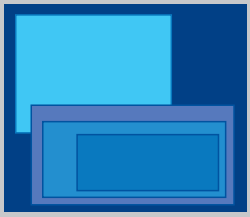


VIAL

VIGO INTERNATIONAL JOURNAL of APPLIED LINGUISTICS

Number 19-2022

Universidade Vigo

Servizo
de Publicacións

VIAL

Vigo International Journal
of Applied Linguistics

Universidade Vigo

VIAL. Vigo International Journal of Applied Linguistics.

VIAL is an international and peer-reviewed journal which is indexed and abstracted in CINDOC (Centro de Información y Documentación Científica), EBSCO, ERIH PLUS, JCR, Web of Knowledge (Arts and Humanities Citation Index, Social Sciences Citation Index), Linguistics Abstracts, Linguistics and Language Behavior Abstracts, Latindex, MLA Directory of Periodicals, PsychoInfo, Scimago Journal Rank Scopus, Ulrich's Database

Webpage: www.webs.uvigo.es/vialjournal/

Printed in Spain - Impreso en España
I.S.S.N.: 1697-0381 - Electronic ISSN: 2660-504X - Depósito Legal: VG-935-2003
Imprime e maqueta: Tórculo Comunicación Gráfica, S.A. Deseño da portada: Elena Gómez.

VIAL

Vigo International Journal of Applied Linguistics

Number 19 - 2022

Contents:

A Matter of Age: Negotiation for Meaning in Child and Adult Interactions

Raúl Azpilicueta-Martínez
Department of Human Sciences and Education
Public University of Navarra (UPNA-NUP)
raul.azpilicueta@unavarra.es

Amparo Lázaro-Ibarrola
Department of Human Sciences and Education
Public University of Navarra (UPNA-NUP)
amparo.lazaro@unavarra.es

Abstract

Findings from task-supported interaction with adult populations have often been transferred to children with little to no modification. When (considerable) differences have been identified, adult and children interactions were analysed while performing different tasks or at different proficiency levels. This article attempts to provide a more reliable comparison by analysing level-matched adults and children performing the exact same task. This study examines the negotiation for meaning (NoM) strategies and their communicative functions in 20 young (age 8-9) children and 14 adults performing an information-gap narrative task with an adult proficient speaker. All participants had Spanish as their L1 and were beginner learners of English as a foreign language (EFL). The results revealed that the adult group produced significantly higher rates in all NoM strategies, with the exception of comprehension checks. However, both populations displayed commonalities in their interactional patterns, with a similar proportional use and functions of the NoM strategies and a clear tendency to imitate elements of their more proficient interlocutor's output via other-repetitions. These findings provide evidence that the age factor at this level of proficiency may have a greater impact on the amount of NoM generated than on the type and function of the NoM strategies used.

Keywords: Interaction; negotiation for meaning; children; adults; EFL.

Resumen

Los hallazgos en estudios de interacción con poblaciones adultas han sido extrapolados a menudo a la población infantil sin apenas modificaciones. Además,

las diferencias encontradas entre niños y adultos provienen de estudios en los que estas poblaciones realizan diferentes tareas o tienen diferentes niveles de competencia lingüística. Este trabajo de investigación pretende ofrecer una comparación más fiable mediante el análisis de interacciones de adultos y niños con el mismo nivel de competencia y realizando la misma tarea. Así, nuestro estudio analiza las estrategias de negociación de significado (NdS) y sus funciones comunicativas en las narraciones de una historia con vacíos de información de 20 aprendices jóvenes (8-9 años) y 14 adultos con un adulto experto. Los participantes compartían el castellano como L1 y tenían un nivel básico en inglés como lengua extranjera. Los resultados muestran que el grupo adulto utilizó más estrategias de NdS, a excepción de las comprobaciones de comprensión. Sin embargo, ambas poblaciones mostraron una proporción similar en el uso y funciones de dichas estrategias y una clara tendencia a imitar elementos de la producción de su interlocutor, más competente lingüísticamente, mediante el uso de repeticiones. Estos hallazgos sugieren que, con bajos niveles de competencia lingüística, el impacto del factor edad puede estar más relacionado con la cantidad de NdS que con el tipo y funciones de sus estrategias.

Palabras clave: Interacción; negociación de significado; niños; adultos; ILE.

1. Introduction

As on frequent occassions in the field of SLA, interaction-based research started with adults (Long, 1983a; Mackey, 1999; Shehadeh, 1999; Varonis & Gass, 1985) and only subsequently did it address children as an object of study (Butler & Zeng, 2014; Mackey & Oliver, 2002; Oliver, 2009). This originally led to the default transfer of general assumptions to the latter population with little or no empirical evidence despite both groups' notable differences (Birdsong, 2005; García Mayo, 2018; García Mayo & García Lecumberri, 2003; Muñoz & Singleton, 2011; Singleton & Ryan, 2004). Coinciding in time with the expansion of foreign language (FL) programmes and content and language integrated learning (CLIL) (Dalton-Puffer, 2011; Sylvén, 2013), a growing body of research warned about the lack of solidity of such transfer, (Mackey & Gass, 2005, following Thompson & Jackson, 1998; Paradis, 2007) and gradually contributed to shed light on the distinct features of children, overlooked in interaction-based research until the late nineties (e.g. Oliver, 1998; Van den Branden, 1997).

Breakthrough studies at the turn of the century outlined the peculiarities in the interactional patterns of children learning English as a second language (ESL) (e.g., Oliver, 1998, 2000, 2002; Oliver & Mackey, 2003; Mackey, Kanganas, et al., 2007) and acted as a catalyst for subsequent research on EFL young learners (YLs)

(Azkarai & García Mayo, 2015; García Mayo & Hidalgo, 2017). Findings in both contexts demonstrated the benefits of NoM for children when interacting with adults and also with peers and allowed researchers to champion the use of interactive tasks as a valuable pedagogical practice with ESL/EFL YLs (Hidalgo, 2019; García Mayo & Imaz Agirre, 2019). However, while children were shown to be able to negotiate for meaning, all studies consistently reported greater amounts of negotiation among adults in aspects such as NoM strategies (e.g., Oliver, 1998, 2000, 2002), reception of negative feedback (Oliver, 2000) provision of feedback and modified output (Oliver, 2000; Mackey, Oliver, et al., 2003), or the use of task-related strategies, although results in aspects such as L1 use seem far less conclusive (Pinter, 2006). However, these claims have been made on the basis of empirical studies carried out either with children or with adults, which implies that the data from each age group were elicited under different conditions.

In fact, very few studies to date have included EFL adult and children populations within the same study and/or performing the same interactive tasks (Pinter, 2006; Azpilicueta-Martínez, 2020). Pinter's (2006) study examined the task solving strategies in both populations, and the results in her analysis provided evidence that adults were more efficient than children in the way they handled the tasks. On the other hand, a study by Lázaro-Ibarrola and Azpilicueta-Martínez (2019b) compared the NoM in child-child and adult-adult interactions and identified important similarities between both age groups as regards both the amount of NoM strategies as well as the functions they served, questioning previous research claims.

Therefore, the few studies comparing adults and children have done so with participants interacting with age-matched peers (i.e., child-child vs. adult-adult). What is more, if we focus specifically on interaction-based studies where children and adults negotiate with an expert speaker of the language (i.e., child-expert vs. adult-expert), research comparing both populations within the same study is simply non-existent. However, it is also known that learner-teacher interaction constitutes a frequent real classroom situation (Tuan & Nhu, 2010), and that it mirrors the most widespread format of oral examinations at the lower levels of the Common European Framework of Reference for Languages (CEFR), facts which highlight the need of more studies analysing interactional patterns in this particular setting even more.

In sum, research comparing children and adult interactions is in dire need of more studies. Specifically, more light needs to be shed on the ways these populations interact with expert speakers of the target language (TL). The present study attempts to address this research niche by specifically analysing the conversational patterns of EFL children and adult beginners at the same level of proficiency while interacting with a proficient speaker of English in an information-gap narrative oral task.

2. Literature review

2.1. The patterns of interaction in EFL children and adults

Interaction-based studies (Long & Porter, 1985; Philp & Tognini, 2009) put forward the benefits of conversational interaction for SLA, since NoM "facilitates language acquisition because it connects input (what learners hear and read); internal learner capacities, particularly selective attention; and output (what learners produce) in productive ways" (Long, 1996: 451-452). The first studies within the interactionist framework operationalised NoM by means of strategies, comprising conversational adjustments (CAs) and different forms of repetition. With the increase of studies on the subject, this canonical inventory has been augmented by additional strategies, namely 'acknowledgements' and supplemented by the inclusion of the communicative functions these strategies serve (see 3.5. Data coding and analysis). A common conclusion arising from these studies is that learners of English as an L2 are able to negotiate for meaning, use the wide range of strategies described above, and overcome communication breakdowns successfully (García Mayo & Imaz Agirre, 2019; Hidalgo, 2019). Also, NoM proved to be affected, among other factors, by the inter-related effects of the proficiciency level of the conversational partners (learner/expert) and the age of the interlocutors (i.e., adults vs. children) (e.g., Mackey & Oliver, 2002).

One of the most consistent findings when comparing conversational patterns in adult and children learners was the relatively higher overall use of strategies of the former (Long, 1983a; Oliver, 1998, 2002, to name but some). However, values for several NoM strategies in dyads made up of native-native (NS-NS) and non-native-native (NNS-NS) children (clarification requests and confirmation checks, respectively, and repetitions in both cases) were higher than those in adult NS-NS dyads (Long, 1983a; Oliver, 1998). This finding was hypothesised to be related to the effect of the 'nativeness or nonnativeness' of the interlocutor in relation to the age of the learners (Oliver, 1998: 377). In other words, the proficiency of a conversational partner affected children and adults in different ways. With the exception of the values for other-repetition, whenever one of the subjects in the adult pairings was a learner (NNS), values in all NoM strategies were substantially higher than those of YLs (Long, 1983a; Oliver, 1998).

In addition to the greater quantity of strategies reported in adult populations, recurrent differences have also been reported when looking at the proportional use of certain types of negotiations. For example, although mean values are still lower for YLs, children have proven to make a wider proportional use of 'self' NoM strategies, and to be barely interested in their partners' understanding if compared with adult

populations (Oliver, 1998, 2009). This feature has been attributed to the egocentricity inherent to YLs, and seems to intensify in the case of younger children, as underpinned by more recent work comparing YLs of different ages (García Mayo & Lázaro-Ibarrola, 2015; Azkarai & Imaz Agirre, 2016; Hidalgo, 2019).

However, as Oliver (1998) acknowledges, some of the aforementioned studies make claims about the differences in NoM between adult and children without having been able to carry out a statistical comparison, given the use of dissimilar methods in data collection, the use of substantially different task types or even number of tasks (e.g., comparison of NoM rates between Long, 1983a, and Oliver, 1998), and the different proficiency levels, age and learning contexts of the interlocutors. Recently, findings in one study specifically comparing EFL children and adult populations in peer interaction while performing the same task under the same conditions did not fully support the assumption that adults negotiate more and that children are not so aware of their interlocutor's needs (Lázaro-Ibarrola & Azpilicueta-Martínez, 2019b). Rather on the contrary, results revealed statistically significant higher rates in child-child interactions on the overall number of CAs as well as in self-repetition, and similarities were reported regarding the type and degree of use of the communicative functions NoM strategies served.

When looking at the proficiency level of the interlocutors researchers have found that, provided a minimum threshold level of proficiency has been attained (Lázaro-Ibarrola & Azpilicueta-Martínez, 2015), the mastery of the TL is inversely proportional to the amount of NoM generated in dyadic interaction. At one end nonnative dyads produced the most NoM, as opposed to native speaker pairings, as the following scale illustrates (H=high; L=low; VL= very low; NS=native speaker; NNS=non-native speaker):

L-L>H-L>H-H>L-NS>H-NS>NS-NS >VL-VL

(Lázaro-Ibarrola & Azpilicueta-Martínez, 2015, adapted from Oliver, 2002)

These findings have been explained by the scarcity of communication breakdowns generated by more proficient learners (Azkarai & Imaz Agirre, 2016; Ellis, 1985; Gass & Varonis, 1985). However, again these rates belong to different studies using different tasks, categories (acknowledgements) and contexts (ESL vs. EFL), and, as mentioned above, the proficiency of the conversational partners appears to affect each age group differently. Recent work by Lázaro-Ibarrola and Azpilicueta-Martínez (2019a) compared the conversational patterns of child-child (L-L) and child-expert (L-near native) interactions. Contrary to the classification above, values for nearly all NoM strategies (except other-repetitions and acknowledgements, the latter being nearly

identical) were higher in the child-child pairings, including statistically significant differences for comprehension checks and self-repetitions.

In sum, both the age (child vs. adult) and the proficiency of the conversational partner (learner-learner or learner-expert/teacher), have proven to affect the patterns of interaction (Hidalgo, 2019; Oliver, 2000). While some research has been conducted comparing children and adults when interacting with peers and has served to challenge and qualify previous findings, to date no study has compared children and adults when interacting with an expert of the TL, under similar conditions and within the same study, thus making a call for more research to further understand the different interactional patterns of each age group.

3. Method

3.1. Research questions

This study examines the role that age plays in the patterns of interaction of EFL children and adults, at an A1 level of proficiency, interacting with a proficient speaker of the TL (teacher) while performing the same task. Our research questions are the following:

1. What are the patterns of interaction of level-matched EFL children and adults performing the same oral task with a proficient speaker of the TL?

2. How do results in this study compare with previous research?

3.2. Participants

The present study examines the patterns of interaction of twenty (20) EFL children and fourteen (14) EFL adults at an A1 level of the CEFR while performing two narrative tasks with a proficient speaker of the TL. The children group included eleven (11) girls and nine (9) boys, with a mean age of 8.5. All children participants were enrolled in Year 3 at a primary state school in a village in northern Spain. At the time of data collection they had been provided with nine (9) weekly EFL 50-minute lessons for six (6) school years, comprising, in turn, five (5) EFL sessions, plus four (4) sessions of subjects taught through the medium of English (CLIL). Prior to the study the children were told that they were going to take part in an English-speaking game, which did not imply any form of test or examination. All parents were duly informed that their children's performances would be used for research purposes exclusively

and would remain anonymous. Due permission was granted by both parents and the school itself.

All 14 adult subjects in this study had enrolled in two A1 courses at two EFL schools for adult learners. The contents and material in those courses met the criteria established by the Common European Framework of Reference for Languages (CEFR) for an A1 level. Their age ranged from 31 to 69, with a mean age of 47. Due permission was granted by all subjects.

The implementation of an oral proficiency placement test with participants was not deemed appropriate for several reasons. For one thing, defining proficiency, a controversial source of debate itself (Pienemann & Johnston, 1987), seems to further complicate when comparing two distinct age groups (Oliver, 2000), as is the case with this study. In addition, familiarity with a given task in EFL has been reported to impinge on a learner's output (Kazemi & Zarei, 2015; Qiu, 2019). Consequently, the proficiency levels of participants were based on a) the children's internal (school) and external (regional Government) assessment and b) the adults' enrolment in the A1 group. The children's internal school assessment was provided by the school teachers and was based on their regular evaluation of the students' performance. The external assessment consisted of a test developed and administered by the Government to all students in the region of Navarra at the end of the 2nd year of primary. As for the adults, no level test was administered, their A1 level was guaranteed by their enrolment in the EFL A1 courses in their schools where students have to demonstrate their level (by means of a test or an official certificate) to be allowed to enrol. Also, at the beginning of each academic year, those students whose language level is deemed higher are encouraged to move up to the A2 level.

The researcher interacting with the subjects was one of the authors in the present study. Although not a NS, his proficiency level was certified by the obtention of the Cambridge Proficiency in English (CPE) exam (level C2 of the CEFR) with a top score in the oral production skill. The researcher did not know the students prior to the study.

3.3. The task

The task consisted of two (2) stories which featured an identical layout, although based on different stories (see Appendices A and B for stories one and two, respectively). In each of them, one participant (a) was provided with a poster story comprising five (5) pictures showing a sequence of numbered events which s/he had to narrate in that specific order. The other participant (b), the 'story-builder' was given a blank poster,

plus eight (8) scattered pictures including the five pictures given to participant (a) plus three (3) additional distractors. In order to favour negotiations, the order in the story provided to participant (a) was not fully predictable, and its distractors were carefully planned with the aim of forcing participants to negotiate for meaning.

The adequacy to generate interaction of the tasks with both populations was piloted before the study, yet participants involved in the piloting were excluded from the final pool of subjects.

3.4. Procedure

Data were collected as follows. On the first story, the researcher narrated the story so that the participants had to arrange the sequence of events correctly and leave the three wrong pictures out. The researcher narrated the story using exactly the same script (see Appendix C), and provided the exact same feedback to all participants depending on their responses. The script included a few warm-up questions, not coded in the study, in order for the learners to get acquainted with the expert speaker and feel more comfortable. An opaque screen was placed between participants in order to maximise verbal communication. The researcher took on the narrator role (a) first with the aim of familiarising the students with the task and and to dispel any feelings of unease before asking them to take on the narrator role. When the participant felt that s/he had finished with the task, the researcher would stand up and double-check the participant's poster. If the pictures were not placed in the same order as (a) he would then point out which picture(s) were wrong, without providing any further information, until the story was solved successfully.

Once the first story was finished the researcher and the subjects swapped roles and interacted again in order to co-construct the second story, that is, the participant, who had acted as the story-builder, would now take on the narrator role, and the researcher, who had the narrator role during the first story, would now become the story-builder. All interactions were filmed and voice-recorded using an integrated webcam and a voice recorder.

3.5. Data coding and analysis

The participants' interactions were directly observed and video-recorded by one of the researchers. Their conversations were then transcribed verbatim, and coded following the NoM categorisation commonly used in interaction-based research (Long, 1983a, 1983b; Oliver, 1998; Pica, 1987), which comprises CAs and repetitions. CAs include, in turn, clarification requests, confirmation checks and comprehension

checks, with the addition of 'acknowledgements' as pointed out in more recent studies (Ducasse, 2010; Galaczi, 2013). Repetitions include instances of self-repetition and other-repetition. We will now describe these elements succinctly and provide examples from the current study.

Clarification requests are used by the listener when they need to make clear what their interlocutor has said (Oliver, 2002: 103). In example 1 below, the child is describing the last image in story 2 (see Appendix B) to the researcher, in which the difference between the right picture and the distractor hinges on the length of the snowman's arms. When the researcher asks about that element the child seems to fail to understand the question and resorts to a clarification request (turn 3), which then triggers a modification in the researcher's wording of the same question (turn 4), and ultimately leads to an understanding of the query and subsequent (successful) response (turn 5):

Example 1 (Child-researcher)

1. Child: The story finish that the children put the, the hat, the nose and the scarf to the snowman and they are, they are happy.
2. Researcher: Ok. Are the snowman's arms long or short now?
3. Child: ***What?*** [Clarification request]
4. Researcher: The snowman's arms: are they long or short?
5. Child: Long.

Confirmation checks are used by the listener when they need to ratify that they have really understood what their interlocutor has said (Oliver, 2002: 103). This is illustrated in example 2 (below; story 2), in which the adult participant requests confirmation (turn 9) of an utterance in which the researcher was co-constructing the story (turns 6 and 8), triggering a confirming response (turn 10), and requesting confirmation again (turn 11). In this case, such persistence appears to be related to the participant failing to see the pictures correctly, suggested by the successful response once she puts her glasses on (turn 15):

Example 2 (Adult-researcher)

1. Adult: Erm... erm... a... y... there is a... a... a man?
2. Researcher: Yes.
3. Adult: Sit in the... *no sé cómo se dice 'banco'* (I don't know how to say 'banco' -'bench').

4. Researcher: He's sitting on a bench.
5. Adult: On a bench.
6. Researcher: Very good, eating a banana.
7. Adult: Eh?
8. Researcher: The man is sitting holding a banana.
9. Adult: ***A banana?*** [Confirmation check]
10. Researcher: Yes.
11. Adult: ***Banana?*** [Confirmation check]
12. Researcher: Yeah....
13. Adult: Banana... The bank.
14. Researcher: Yes.
15. Adult: Ah!!! *Vale, espera* ('Ok, hang on') (Adult puts on glasses). Yes! Eating a banana, yes, yes, yes.

Comprehension checks are performed by the speaker - as opposed to the previous two CAs - to ensure that the preceding utterance has been fully understood by the listener (Oliver, 2002):

Example 3 (Adult-researcher)

1. Adult: The, the girl and the, and the boy make, make the ball.
2. Researcher: But, who is making the bigger ball: the girl or the boy?
3. Adult: The girl, the girl.
4. Researcher: The girl, yes? Ok, thank you.
5. Adult: Erm... ***Number three?*** [Comprehension check]
6. Researcher: Yes.
7. Adult: They make a... a.... (pause) d... erm... doll? Doll... snow...dolls? They... make two... two... two, two, two eyes, and a.... erm... *Es que no sé cómo se dice 'muñeco'*.(I don't know how to say 'muñeco' – 'snowman')

In example 3 (story 2) there has been some succesful NoM in the preceding utterances (turns 1-4), and then, the adult participant asks whether she can start describing the following picture (turn 5), i.e., implicitly asking whether the previous

utterance (turn 3) has been fully understood by her interlocutor so that they can move on in the task. While this instance does not constitute a word-for-word sample of the types of comprehension checks pointed out by Oliver ('either tag questions, repetition with rising intonation, or questions such as 'Do you understand?'; Oliver 2002: 103), it does perform the role of a comprehension check in this specific task. In fact, researchers have already warned about the risk of overrestricting the coding of CAs to the literal examples provided in the literature, since that might lead to "inaccurate rates not reflecting the actual use of conversational adjustments displayed by participants" (Lázaro-Ibarrola & Azpilicueta-Martínez, 2019b: 20).

As opposed to comprehension checks, acknowledgements are used by the listener to confirm to the speaker that the previous utterance has been understood properly (Hidalgo, 2019). This may be seen in example 4:

Example 4 (Child-researcher)

1. Researcher: There are two girls having fun. They're playing with a doll. The doll is in the cot. They're celebrating a birthday party.
2. Child: ***Ah… Number two.*** [Acknowledgement]
3. Researcher: The girls are jumping happily because mum has brought them a cat.
4. Child: ***Yes. Number four.*** [Acknowledgement]
5. Researcher: The girls are now looking for the cat. They cannot seem to find it!
6. Child: Yes.

Similar to example 3, the researcher has described picture 1 (turn 1) and the child lets him know (turn 2) that he can move on to the next picture, thus intimating that he had fully understood the previous picture.

Repetitions are referred to as an interlocutor's repetition of lexical items from preceding utterances within five speaking turns (Pica & Doughty, 1985) and they may involve repeated instances of their own production (self-repetition) or their partner's production (other repetition). It is appropriate to note that whenever a form of repetition overlapped with any other CA in the present study, they were only coded as a CA, since, as Oliver (1998) warns, including them in both categories would lead to a distortion of the total number of strategies. This may be clearly noted in the following extract from example 2 above.

Extract from example 2 (Adult-researcher)

8. Researcher: The man is sitting holding a banana.
9. Student A: ***A banana?*** [Confirmation check]
10. Researcher: Yes.
11. Student A: ***Banana?*** [Confirmation check]
12. Researcher: Yeah....
    Student A: Banana... The bank. [~~Other-repetition: partial~~] [Confirmation check]

All the strategies above were classified according to the communicative functions they performed. Thus, NoM strategies were classified according to whether they were used to a) repair a communication problem, including clarification requests and confirmation checks, b) confirm communication, which comprises acknowledgements and other repetitions, and c) prevent breakdowns in communication, which includes comprehension checks and self-repetitions. It is worth noting that while the two first functions (repairing and confirming communication) are listener-generated, the last one (preventing communication breakdowns) is the only function carried out by the speaker.

The analysis of each strategy was carried out considering the total number of strategies the participants used divided by the total number of utterances. One of the researchers coded all the transcriptions and the other researcher coded 20% of the whole dataset independently. Inter-rater reliability reached 95%. The remaining discrepancies were solved individually on a case-by-case basis.

All values were entered into SPSS (version 24) and a normality test was run, which yielded a non-normal distribution. Therefore, the nonparametric Mann-Whitney U-test was used in all statistical analyses in this study. Significance level was fixed at $p = 0.05$.

4. Results and discussion

The first part of this section presents the results of the comparison of NoM strategies and the functions they serve in level-matched EFL children and adults performing the same oral task with an expert (research question 1), while the second part compares findings in the present study with previous research on the subject (research question 2).

Results comprising values, percentages and statistical findings for NoM strategies in both age groups are presented in Tables 1 through 3. Table 1 displays the results for CAs and acknowledgements. Adult-expert dyads generated significantly higher NoM rates than child-expert interaction (u= 24.500, p = < .001[b]), indicating a significant effect for the age variable, corroborating findings in Oliver (1998) and Long (1983a), which revealed that, whenever there is a NNS adult present in the dyad, NoM rates are boosted:

Table 1. Child-expert vs. adult-expert: Conversational adjustments and acknowledgements

		Child- expert		**Adult-expert**		
		Average time: 3 mins 34 secs		Average time: 5 mins 55 secs		
		Average number of utterances: 19.78 (Total 791)		Average number of utterances: 41.43 (Total 1160)		
		Raw	%	Raw	%	**Statistics**
Conversational Adjustments	Clarification Requests	17	2.15%	61	5.26%	(u= 25.500. p < .001[b]) (sig. higher in adult interaction)
	Confirmation Checks	3	0.38%	40	3.45%	(u= 39.000. p < .001[b]) (sig. higher in adult interaction)
	Comprehension Checks	0	0%	1	0.09%	(u= 130.000. p = 0.743[b]) (non-significant)
	Total	**20**	**2.53%**	**102**	**8.79%**	(u= 16.500. p = < .001[b]) (sig. higher in adult interaction)
Acknowledgements		47	5.94%	72	6.21%	(u= 16.500. p = < .001[b]) (sig. higher in adult interaction)
Total number of strategies		**67**	**8.47%**	**174**	**15%**	(u= 24.500. p < .001[b]) (sig. higher in adult interaction)

It may be noted at this point that the higher NoM values present in the adult-expert interactions are associated with a substantially larger amount of language produced, since the older learners produced an average of 41.43 utterances per participant, whereas children produced less than half as many: 19.78. This phenomenon was also reflected in the amount of time spent performing the task (five minutes and 55 seconds for the adults, in contrast to three minutes and 34 seconds for the chidren).

Equally interesting, both age groups displayed very similar conversational patterns, since the proportional use of each conversational adjustment between both groups was strikingly similar: both populations resorted primarily to acknowledgements, followed by clarification requests, confirmation checks, and, lastly, comprehension checks. Examining each strategy in isolation, it may be noted that adults more than doubled the percentage of clarification requests with respect to children, yielding statistically significant differences (u= 25.500, p < .001[b]). In the case of confirmation checks, the gap increased nearly tenfold, and differences were significant (u= 39.000, p < .001[b]). Values for comprehension checks were nearly inexistent (indeed so in the case of children), and differences between both groups non significant (u= 130.000, p = 0.743[b]). While the scarcity of comprehension checks in children populations has already been accounted for in the literature due to the egocentricity inherent to children (e.g., Azkarai & García Mayo, 2016: 10; Oliver, 1998: 377), the extremely low rate in the adult group might be related to their assuming that their more proficient interlocutor would understand them at all times.

Acknowledgements, as mentioned above, constituted the single most frequently used conversational adjustment by both populations, yet adult-expert interaction displayed a significantly higher number of instances of this strategy than children-expert interaction (u= 77.500, p = 0.027 [b]).

Table 2 shows the values, percentages and results of statistical analyses for repetitions. Overall, the adult group produced significantly more instances of repetitions than the children in the study (u= 1.000, p < .001[b]). This is further borne out by differences in both self-repetition (u= 39.500, p < .001[b]) and other-repetition (u= 5.000, p < .001[b]). Once more, however, both groups feature a similar proportional use of each of these elements, with a strong preference for other-repetition (19.83% and 11.76% for adults and children respectively) over self-repetition (12.67% and 5.94%).

Table 2. Child-expert vs. adult-expert: Repetitions

		Child- expert		**Adult-expert**		
		Total number of utterances: 791		Total number of utterances: 1160		
		Raw	%	Raw	%	**Statistics**
Repetitions	Self-repetition	47	5.94%	147	12.67%	(u= 39.500, p < .001[b]) (sig. higher in adult interaction)
	Other-repetition	93	11.76%	230	19.83%	(u= 5.000, p < .001[b]) (sig. higher in adult interaction)
	Total	**140**	**17.70%**	**377**	**32.5%**	(u= 1.000, p < .001[b]) (sig. higher in adult interaction)

It is immediately apparent that, beyond the substantial differences in the above-mentioned amount of NoM produced lies a common denominator: children-expert interactions resemble a scaled-down replica of the adult-expert conversational patterns, which might be explained by their common proficiency level. Although the adult group made a much more extensive use of the strategies, the proportional use of each of the elements analysed was very similar.

Let us now examine the communicative functions that the elements in Tables 1 and 2 serve in interaction (Table 3).

Table 3. Child-expert vs. adult-expert: Communicative functions

		Child-expert		**Adult-expert**		
		Utterances: 791 Total strategies 207		Utterances: 1160 Total strategies 551		
		Raw	%	Raw	%	**Statistics**
Communicative Function	Repair (clarification requests & confirmation checks)	47	22.70%	144	8.71%	(u= 19.000, p < .001[b]) (sig. higher in adult interaction)
	Confirm (other-repetitions & acknowledgements)	140	67.63%	306	26.04%	(u= 6.500, p < .001[b]) (sig. higher in adult interaction)
	Prevent (self-repetitions & comprehension checks)	20	9.66%	101	12.76%	(u= 39.000, p < .001[b]) (sig. higher in adult interaction)

Since each of the categories above is composed of the same elements in Tables 1 and 2, in which adults showed significantly higher rates in all but one strategy (comprehension checks), it is expected that the YLs made use of these functions to a significantly lower degree. Perhaps more importantly, the data above provide compelling evidence that adults and children used NoM in order to achieve the same communicative functions, with a clear preference for confirming communication, followed by repair, with the prevention of communication breakdowns in last place.

Research question 2 intended to compare the results in the present study with those from previous research. In order to do so, the following studies analysing NoM in children and adult populations included in the literature review have been selected (Table 4):

(Children)

1. Oliver (1998): results from ESL NNS-NS child-child interactions, since that type of dyad is closest to the type of proficiency pairing (NNS-expert) in the present study.

2. Lázaro-Ibarrola & Azpilicueta Martínez (2019b): results from same-level EFL NNS-NNS child-child interactions using tasks with identical layout, yet different content, to the ones in the present study.

(Adults)

1. Long (1983a): results from ESL NNS-NS adults.

2. Lázaro-Ibarrola & Azpilicueta Martínez (2019b): results from same-level EFL NNS-NNS adult-adult interactions using tasks with identical layout, yet different content, to the ones in the present study.

Table 4. NoM in children and adult studies

		Oliver (1998)	**LI-AM (2019b)**	**Pres. Study**	**Long (1983a)**	**LI-AM (2019b)**	**Pres. Study**
		ESL	**EFL**	**EFL**	**ESL**	**EFL**	**EFL**
		CHILDREN			ADULTS		
		NNS-NS	NNS-NNS	NNS-expert	NNS-NS	NNS-NNS	NNS-expert
		Child-child	**Child-child**	**Child-expert**	**Adult-adult**	**Adult-adult**	**Adult-expert**
Conversational Adjustments (CAs)	Clarification Requests	3.97	3.69	2.15	10.35	1.44	5.26
	Confirmation Checks	6.44	1.02	0.38	18.15	1.75	3.45
	Comprehension Checks	0.41	1.66	0	18.15	1.52	0.09
	Acknowledgements	-	8.41	5.94	-	7.06	6.21
	Total CAs	**10.82**	**14.78**	**8.47**	**46.65**	**11.77**	**15.01**
Self-repetition		21.66	29.81	5.94	41.06	27.09	12.67
Other-repetition		22.91	8.54	11.76	15.09	12.52	19.83
Total repetition		**44.57**	**38.35**	**17.7**	**56.15**	**39.61**	**32.5**
Total NoM strategies		**55.39**	**53.13**	**26.17**	**102.8**	**51.38**	**47.51**

First and foremost the overall similarities in the conversational patterns in the present study, which portray child-expert interactions as a scaled-down replica of adult-expert interactions, are consistent with initial findings comparing ESL NNS-NS child-child (Oliver, 1998) and NNS-NS adult-adult (Long, 1983a) interactions. These results support the notion that not only ESL, but also EFL dyads including a NNS adult interacting with a NS/expert trigger the highest amount of NoM.

However, these proportionally similar conversational patterns were not found in Lázaro-Ibarrola & Azpilicueta Martínez (2019b), which compared EFL NNS-NNS child-child and NNS-NNS adult-adult interaction. What is more, the patterns in their study revealed significantly higher NoM rates for the YLs in the overall CAs and self-repetitions percentages, pointing at a clear influence of the interlocutor factor on the type and amount of NoM. In fact, the child-expert combination in the current study reduces overall NoM rates to less than half if compared to the NNS-NNS child-child interaction in Lázaro-Ibarrola & Azpilicueta Martínez (2019b), i.e., from 53.13% to 26.17%, and the same applies for CAs: from 14.78% to 8.47%. In other words, YLs' NoM seems to dwindle when interaction takes place with an adult expert.

This phenomenon does not occur with the NNS-NNS adult group, and, although there is a decrease in overall NoM rates (from 51.38% in Lázaro-Ibarrola & Azpilicueta Martínez, 2019b, to 47.51% in the present study), this is far less substantial. In fact, the tendency reverses if we focus on CAs in isolation, since the adult-expert combination experienced an increase if compared with the adult-adult interactions in Lázaro-Ibarrola & Azpilicueta Martínez (2019b) (from 11.77% to 15.01%).

These opposing results might be related to the interlocutor factor, and suggest that YLs might be more sensitive than adults to differences in the age and perceived role of their conversational partners. The children in the study may have adopted a more submissive role when interacting with an expert they may have associated with a 'teacher role', an observation already pointed out in the literature (Scarcella & Higa, 1981), an influence to which the adults might have been more oblivious. The fact that the expert speaker in the dyad was an adult might have also had a positive impact on the NoM displayed by the adult group. Interestingly, both populations seem to have a tendency to imitate the expert's speech (note the increase in other-repetition rates) if compared to peer interaction.

Secondly, the amount of NoM in all strategies is clearly lower than either of the ESL studies compared in both populations, concurring with previous research comparing NoM in ESL and EFL populations (e.g., García Mayo & Hidalgo, 2017; Oliver, 1998), yet seems to be past the 'red line' or minimum threshold level in Lázaro-Ibarrola & Azpilicueta Martínez (2015). However, it is important to note at this point that data

coding in Long's study (1983a) with NNS-NS adult learners might have been different to that in the rest of studies subject to scrutiny here. This may be noted in the fact that the value for 'Total NoM strategies' in Long's study exceeds 100%. As mentioned in section 3.5. (Data coding and analysis), this was avoided in the rest of the studies analysed, as whenever a CA and a form of repetition overlapped, such strategy was only coded as a CA in order to avoid a misrepresentation of the percentages, following Oliver (1998).

5. Conclusion, limitations and pedagogical implications

The present study has investigated the NoM present in the interactions of children and adults with a proficient speaker of the TL. The first research question intended to compare the interactional patterns of these two populations. An analysis of the NoM strategies and functions displayed by the participants has revealed that the adult group were able to negotiate significantly more than children in all NoM strategies, with the exception of comprehension checks, although these were also higher in the adult group. These results underpin the notion that, when comparing low level EFL learners interacting with an adult expert, older students are at an advantage in terms of their outright production of NoM if compared with YLs, who might appear to take on a more submissive role. As the language level was similar in both populations the main reason for this difference seems to be the young age of the learners, who are not mature enough to play a leading role in most activities and, linguistically speaking, cannot use NoM strategies as much as adults do. More research would be needed in order to ascertain whether a child expert partner would yield different results with these two EFL age groups. Likewise, it would be interesting to know if similar findings could be revealed when students negotiate in their own L1s.

Results have also exposed strikingly similar conversational patterns in the proportional use of NoM strategies between both age groups, i.e., children use the same types of strategies, and for the same purposes as adults. This finding might be explained by the proficiency level exerting a homogeneising effect on the types and functions of NoM present in beginner EFL learners' interactions with an expert.

The second research question intended to examine how the results in the present study compared to the existing research. Firstly, results have corroborated previous findings on ESL child and adult learners, and indicate that dyads including an expert and a NNs adult trigger the highest NoM rates of all proficiency and age combinations. Secondly, the age-NoM relation of EFL NNS-expert interactions is similar to that in ESL NNS-NS interactions from previous research. In other words, children produce the same types of NoM strategies as adults to a lesser, yet proportional, extent.

However, if we compare the learner-expert combination in the present study with EFL peer-peer interaction in existing research using the same task, with age and level-matched pairs, we are faced with opposing interactional patterns. Children in peer interaction negotiate for meaning significantly more than adults in that form of interaction, and significantly more than children interacting with an expert, a common form of interaction (student-teacher) in EFL classes worlwide. Peer-peer interactions see children's NoM rates increase in all strategies (CAs and repetitions), with the exception of other-repetition. This exception could be explained by their perception of their conversational partner as an 'expert' leading to them trying to imitate their speech, at the same time they might take on a more submissive role, thus triggering fewer instances of NoM.

All in all, this study questions the notion of a clear-cut inversely proportional relation between proficiency in the TL and amount of NoM with EFL learners, and suggests it is more complex and dynamic, one in which NoM is mediated by age, type of learning (ESL vs EFL) and interlocutor status, notwithstanding other variables out of the scope of the present study, such as task typology. Much more research is needed in order to ascertain the type of influence each of these elements exert on each population.

The present study contains several methodological limitations that should be addressed in further research. First, the tasks we used were novel and had not been extensively used in previous research. Also, the fact that the researcher specifically pointed out which picture(s) were misplaced when both posters did not match constitutes a limitation, because, on occasions, the participants would then simply swap the position of two pictures without negotiating for meaning any further. If the researcher had simply let them know that one (or several) picture(s) were misplaced, this might have forced participants to double-check the whole story following a (probably) more meticulous approach, and, possibly, negotiating for meaning more. This was done for time constraints primarily. On the other hand, the story itself could have been more suitable for one of our populations, for instance, perhaps the adults found it too simple, which could, in turn, have affected their interaction.

Findings in the study point towards the pedagogical implication of implementing in-class peer-peer interaction activities as a regular classrooom practice with low-level EFL YLs, as opposed to more traditional student-teacher interactions, since that age group seems to benefit from more opportunities for NoM in that type of pairing than adults, who appear to negotiate for meaning more when interacting with a proficient speaker of the TL. However, the results also highlight the potential of learner-expert combinations as a valuable means to promote the activation of specific structures and lexicon in the learners' production via other-repetitions, since the low proficiency

learners (including YLs) in the study tended to imitate more of their partners' speech as a perceived expert speaker of the TL. Consequently, different dyad combinations would appear to yield different learning potentialities (i.e., imitation of particular structures/lexicon versus promoting negotiation of meaning). Finally, the study ultimately encourages language practitioners to become good models of language use, that is, to achieve high levels of proficiency in the language they teach, since both children and adults seemed to imitate their conversational partner's speech when they perceived him/her as an 'expert', a phenomenon which might be frequent in learner-teacher interactions worldwide.

6. Acknowledgements

The present study paper was funded by the Ministerio de Economía y Competitividad (Funder ID: http://dx.doi.org/10.13039/501100003329, Grant Number: FFI2012-32212). We gratefully acknowledge the assistance of the staff at CPEIP Lorenzo Goicoa, UNED Pamplona-Iruña and EOIDNA, in particular to Juana Mª Urbiola and Olga Zamanillo. Finally, we are also thankful to Israel Azpilicueta for letting us use his drawings for the tasks in this study.

7. References

Alegría de la Colina, A., & García Mayo, M.P. (2009). Oral interaction in task-based EFL learning: The use of the L1 as a cognitive tool. *International Review of Applied Linguistics in Language Teaching*, *47*(3-4), 325-345.

Azkarai, A., & García Mayo, M.P. (2015). Task-modality and L1 use in EFL oral interaction. *Language Teaching Research*, *19*(5), 550-571.

Azkarai, A., & Imaz Agirre, A. (2016). Negotiation of meaning strategies in child EFL mainstream and CLIL settings. *Tesol Quarterly*, *50*(4), 844-870.

Azkarai, A., & Oliver, R. (2019). Negative feedback on task repetition: ESL vs. EFL child settings. *The Language Learning Journal*, *47*(3), 269-280.

Azpilicueta-Martínez, R. (2020). Verbal evidence of task-related strategies in EFL: Children and adult interactions. *International Journal of English Studies*, *20*(3), 1-28.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. DeGroot (Eds.), *Handbook of Bilingualism: Psycholinguistic perspectives* (pp. 109-127). New York: Oxford University Press.

Butler, Y. G., & W. Zeng. (2014). Young foreign language learners' interactions during task-based paired assessment. *Language Assessment Quarterly, 11*(1), 45-75.

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182-204.

Ellis, R. (1985). Teacher-pupil interaction in second language development. In S. M. Gass, & C. G. Madden (Eds.), *Input in Second Language Acquisition* (pp. 69-85). Rowley, MA: Newbury House.

García Mayo, M.P., (2018). Child task-based interaction in EFL settings: Research and challenges. *International Journal of English Studies*, *18*(2), 119-143.

García Mayo, M.P., & García Lecumberri, M.L. (2003). *Age and the Acquisition of English as a Foreign Language*. Clevedon: Multilingual Matters.

García Mayo, M.P., & Lázaro-Ibarrola, A. (2015). Do children negotiate for meaning in task-based interaction? Evidence from CLIL and EFL settings. *System*, *54*, 40-54.

García Mayo, M.P., & Hidalgo, M. (2017). L1 use among young EFL mainstream and CLIL learners in task-supported interaction. *System*, *67*, 132-145.

García Mayo, M.P., & Imaz Agirre, A. (2019). Task modality and pair formation method: Their impact on patterns of interaction and LREs among EFL primary school children. *System*, *80*, 165-175.

Gass, S. M., & Varonis, E. M. (1985). Task variation and nonnative/nonnative negotiation of meaning. In S. M. Gass, & C. G. Madden (Eds.), *Input in Second Language Acquisition* (pp. 149-161). Rowley, MA: Newbury House.

Kazemi, S. A., & Zarei, L. (2015). The efficacy of topic familiarity on oral presentation: Extensive speaking assessment task of Iranian EFL learners in TBLT. *International Journal of Applied Linguistics and English Literature*, *4*(3), 93-97.

Lázaro-Ibarrola, A., & Azpilicueta-Martínez, R. (2015). Investigating negotiation of meaning in EFL children with very low levels of proficiency. *International Journal of English Studies*, *15*(1), 1-21.

Lázaro-Ibarrola, A., & Azpilicueta-Martínez (2019a). Spotting the differences between child-child and child-adult interactions: Evidence from Spanish EFL learners at low levels of proficiency. In J. Rokita-Jaśkow and M. Ellis (Eds.), *Early Instructed Second Language Acquisition: Pathways to Competence* (pp. 80-106). Clevedon: *Multilingual Matters*.

Lázaro-Ibarrola, A., & Azpilicueta-Martinez, R. (2019b). Negotiation of meaning in child-child vs. adult-adult interactions: Evidence from low proficiency EFL learners. *International Review of Applied Linguistics in Language Teaching*, (ahead-of-print).

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of Second Language Acquisition* (pp. 413-468). San Diego: Academic Press.

Long, M. H., & Porter, P. A. (1985). Group work, interlanguage talk and second language acquisition. *TESOL Quarterly, 19*(2), 207-227.

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition, 21*(4), 557-587.

Mackey, A., & Oliver, R. (2002). Interactional feedback and children's L2 development. *System, 30*(4), 459-477.

Mackey, A., Oliver, R., & Leeman, J. (2003). Interactional input and the incorporation of feedback: An exploration of NS–NNS and NNS–NNS adult and child dyads. *Language Learning, 53*(1), 35-66.

Mackey, A., & Gass, S. (2005). *Second Language Research. Methodology and Design.* Mahwah: Erlbaum.

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching, 44*(1), 1-35.

Oliver, R. (1998). Negotiation of meaning in child interactions. *The Modern Language Journal, 82*(3), 372-386.

Oliver, R. (2000). Age differences in negotiation and feedback in classroom and pairwork. *Language Learning, 50*(1), 119-151.

Oliver, R. (2002). The patterns of negotiation for meaning in child interactions. *The Modern Language Journal, 86*(1), 97-111.

Oliver, R. (2009). How young is too young? Investigating negotiation of meaning and feedback in children aged five to seven years. In A. Mackey & C. Polio (Eds.), *Multiple Perspectives on Interaction. Second Language Research in Honor of Susan M. Gass* (pp. 135-156). New York: Routledge.

Oliver, R., & Mackey, A. (2003). Interactional context and feedback in child ESL classrooms. *The Modern Language Journal, 87*(4), 519-533.

Paradis, J. (2007). Second language acquisition in childhood. In E. Hoff, & M. Shatz (Eds.) *Blackwell Handbook of Language Development* (pp. 387-406). Malden MA: Blackwell.

Philp, J., & Tognini, R. (2009). Language acquisition in foreign language contexts and the differential benefits of interaction. *International Review of Applied Linguistics, 47*, 245-266.

Pienemann, M., & Johnston, M. (1987). *A Predictive Framework of SLA.* Unpublished manuscript. University of Sydney and AMES NSW, Australia.

Pinter, A. (2006). Verbal evidence of task related strategies: Child versus adult interactions. *System*, *34*(4), 615-630.

Pinter, A. (2007). Some benefits of peer-peer interaction: 10 year-old children practicing with a communicative task. *Language Teaching Research, 11*(2), 189-207.

Qiu, X. (2020). Functions of oral monologic tasks: Effects of topic familiarity on L2 speaking performance. *Language Teaching Research, 24*(6), 745-764.

Scarcella, R.C., & Higa, C. (1981). Input, negotiation, and age differences in second language acquisition 1. *Language Learning*, *31*(2), 409-434.

Shehadeh, A. (1999). Non-native speakers' production of modified comprehensible output and second language learning. *Language Learning*, *49*(4), 627-675.

Singleton, D., & Ryan, L. (Eds.). (2004). *Language Acquisition: The Age Factor.* Clevedon: Multilingual Matters.

Sylvén, L. K. (2013). CLIL in Sweden–why does it not work? A metaperspective on CLIL across contexts in Europe. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 301-320.

Thompson, R.A., & Jackson, S. (1998). Ethical dimensions of child memory research. *Applied cognitive psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *12*(3), 218-224.

Tuan, L. T., & Nhu, N. T. K. (2010). Theoretical review on oral interaction in EFL classrooms. *Studies in Literature and Language*, *1*(4), 29-48.

Van den Branden, K. (1997). Effects of negotiation on language learners' output. *Language Learning, 47*(4), 589-636.

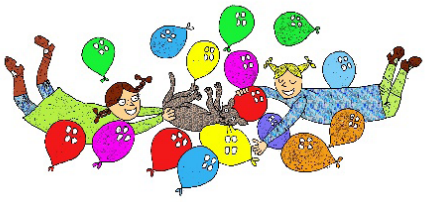
Varonis, E. M., & Gass, S. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics*, *6*(1), 71-90.

Appendices

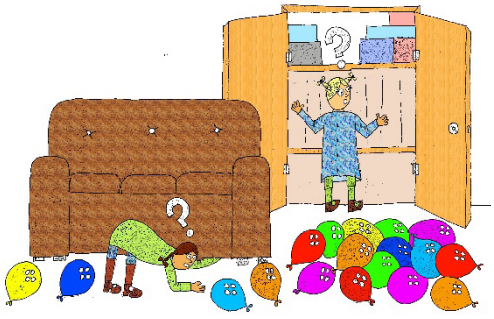
Appendix A – Story 1: *The Birthday Party*

1

2



3

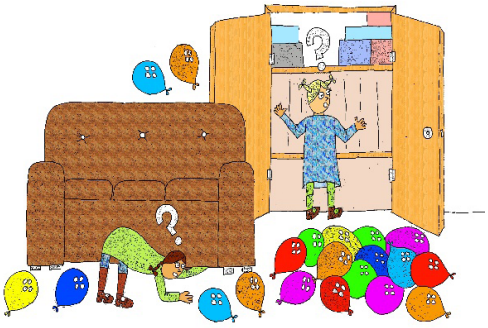


4

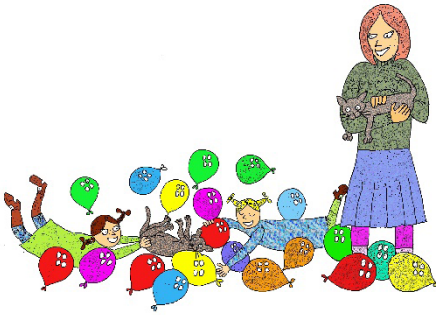
5

Distractors (story 1)

1



2



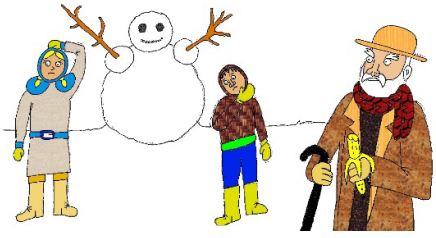
3

Appendix B – Story 2: *The Snowman*

1

2

3



4

5

Distractors (story 2)

1

2

3

Appendix C – Script.

1. Warm-up, ice-breaking conversation

Input for adults	**Backup language**	**Feedback**
Hello/good morning/ good afternoon/evening		Ok/Well done/ Alright/Great/ Excellent
How are you/doing?	Are you ok? Are you alright? *(thumbs up sign)*	
What's/are your name (s)?	My name is.... You are....? *(pointing at self)*	
Where are you from?	Do you live in Pamplona?	

Input for children	**Backup language**	**Feedback**
Hello/good morning/ good afternoon /evening		Ok/Well done/Alright/ Great /Excellent
How are you/doing?	Are you ok? Are you alright? *(thumbs up sign)*	
What's your name?	Are you (saying subject's name)...?	
How old are you?	Are you 7, 8 or 9?	
Where do you live?	Do you live in Pamplona/ Navarra...?	
Have you got any brothers or sisters?	Have you got one brother/ sister?	
What is your favourite sport/hobby?	Do you like football/tennis?	

2. Script for the stories

Now then (Subject A's name), here's a story in pictures. (Subject A's name), the story is mixed up, it's not in order. I'm going to tell you the story. You must listen to me and put the story in order next to these numbers – 1, 2, 3, 4, 5. Do you understand what you have to do?

Let's do an example. I have the story and you put these pictures in order next to the numbers.

Researcher: *Picture 1: There are two girls having fun. They're playing with a doll. The doll is in the cot. They're celebrating a birthday party.*

Researcher: *Picture 2: The girls are jumping happily because mum has brought them a cat.*

Researcher: *Picture 3: The children are playing with the cat on the floor. The place is full of balloons!*

Researcher: *Picture 4: The girls are now looking for the cat. They cannot seem to find it!*

Researcher: *Picture 5: Oh look at it! It was sleeping on the cot all the time!*

2.- *It's your turn now. Now then (Subject A's name), here's a story in pictures. I have the same story but the story is mixed up, it's not in order. (Subject A's name): you have to tell me the story but you must not show me the pictures.*

STORY 2					
Picture 1 - Girl and boy playing in the snow in the park. - Old man on bench in background with hat, scarf, eating a banana.	**Picture 2** - They start to make a snowman, rolling big and small snowballs.	**Picture 3** - Body, head and mouth made - girl putting on stones for eyes - Boy putting in sticks for arms.	**Picture 4** - Snowman nearly finished but missing nose, hat, scarf. - Children wondering. - Old man looking at snowman.	**Picture 5** - Smiling old man and happy subjects looking at snowman. - Snowman wearing hat and scarf from old man, plus banana for nose.	
Backup questions					Feedback (choose from)
Picture 1 *How many subjects are there?* *Are they boys or girls?* *What are they doing?* *Are they running?* *What's the weather like?* *Is it snowy?* *What is the old man doing?* *Is he eating a banana?*	**Picture 2** *What are the subjects doing?* *Are they making snowballs?* *Are the snowballs different?* *What is different?* *Is this a big snowball and this a small snowball?*	**Picture 3** *What parts of the snowman are finished?* *Does the snowman have a body/head/mouth?* *What is the girl doing?* *Is she putting on the eyes?* *What is the boy doing?* *Is he putting on the arms?*	**Picture 4** *Is the snowman finished?* *What is missing?* *Does the snowman have a nose?* *What are the subjects doing?* *Are they looking for a nose?* *What is the old man doing?* *Is he looking at the snowman?*	**Picture 5** *What are the subjects doing?* *Are they happy?* *Why are they happy?* *Is the snowman finished now?* *What did they use for a nose?* *What else is on the snowman?* *Is the snowman wearing a hat and a scarf?*	*Well done/ That's right/ Ok/Brilliant/ Excellent/ Good*

Assessing the impact of the Lexical Approach on EFL perceived oral proficiency: What is the role of formulaic sequences?

Marco Cancino
Department of English
Andres Bello University, Chile
marco.cancino@unab.cl

Jonathan Iturrieta
Department of English
Andres Bello University, Chile
j.iturrietaaguero@uandresbello.edu

Abstract

The Lexical Approach (LA) is a pedagogical method that emphasizes authentic language and learner exposure to co-occurring lexical units. The approach has garnered renewed interest over the last years due to its focus on the frequency of multi-word form-meaning distributions, which is in line with usage-based (UB) approaches to language acquisition. Thus, the present study sought to assess the impact of the LA on perceived oral proficiency and formulaic sequence use. To this end, 38 English as a Foreign Language adult learners at a language institute were divided into two groups, one of which was exposed to 38 teaching hours of instruction based on LA principles. Both groups were asked to perform two oral tasks that were evaluated by three judges in terms of oral proficiency and number of formulaic sequences used. Results revealed that the LA group outperformed the control group in perceived overall oral proficiency scores and in the number of formulaic sequences used, and that there were positive moderate correlations between these measures. Pedagogical implications emphasizing the importance of exposing learners to lexical chunks by means of authentic materials are discussed, as well as their relevance in EFL contexts that focus on the short-term achievement of grammatical accuracy.

Keywords: lexical approach; collocations; formulaic sequences; oral proficiency; EFL learning.

Resumen

El Enfoque Léxico (LA) es un método pedagógico que pone el énfasis en el lenguaje auténtico y la exposición del alumno a unidades léxicas concurrentes. Este

enfoque ha ganado un interés renovado en los últimos años debido a su atención en la frecuencia de asociaciones de forma-significado de múltiples palabras, lo que está en línea con enfoques basados en el uso (*usage-based*) para la adquisición del lenguaje. Por lo tanto, el presente estudio buscó evaluar el impacto del LA en la habilidad oral y el uso de las secuencias de fórmula. Con este fin, 38 estudiantes adultos de inglés como lengua extranjera (EFL) en un centro de idiomas se dividieron en dos grupos, uno de los cuales fue expuesto a 38 horas de enseñanza de instrucción basada en los principios de LA. Se pidió a ambos grupos que realizaran dos tareas orales que fueron evaluadas por tres jueces en términos de competencia oral y número de secuencias de fórmulas utilizadas. Los resultados revelaron que el grupo de LA superó al grupo de control en competencia oral general percibida y en el número de secuencias de fórmula utilizadas, y que hubo correlaciones moderadas positivas entre estas medidas. Se discuten las implicaciones pedagógicas que enfatizan la importancia de exponer a los estudiantes a fragmentos léxicos por medio de materiales auténticos, así como su relevancia en contextos de inglés como lengua extranjera que se centran en la búsqueda de la corrección gramatical a corto plazo.

Palabras clave: enfoque léxico; colocaciones; secuencias de fórmula; competencia oral; aprendizaje del inglés como lengua extranjera.

1. Introduction

The lexical approach (henceforth LA) is a pedagogical method introduced by Lewis (1993) that regards language learning as occurring through learners' repeated exposure to authentic, real-life English in the form of multi-word items being taught as wholes. This approach proposes that the internalization of idiomatic and prefabricated structures to understand and produce language is crucial for successful communication. The LA has garnered renewed interest due to its focus on the frequency of form-meaning distributions, an aspect that is shared by usage-based (UB) approaches to language acquisition. The main tenets of UB theories are that language learning takes place incidentally and implicitly (Madlener, 2018), and that this process is influenced by the frequency of form-meaning associations in the input found in social interactions (Ellis, 2013). Thus, UB approaches to language learning underscore the impact of frequency distributions, salience, and contingency of form-meaning associations (i.e., collocations) on second language development (Tyler et al., 2018). The frequency of repetitions can help commit certain form-meaning structures to memory, increase the abstract representations of such constructions, and enhance their productive use (Ellis, 2013; Madlener, 2018). Being part of such "usage events" in repeated, contextualized social interactions is central to usage-based theories (Tyler & Ortega, 2018). In line with this, pedagogical methods such as the LA have underscored

the idea that contextualized form-meaning distributions based on repetitions can support second language development.

Lexical chunks are form-meaning distributions that play a significant role in language learning (Wood, 2010). These constructions are also referred to as ready-made chunks, multi-word units, and formulaic sequences (Hou et al., 2016). They are fairly common in native and native-like spoken language and thus are seen as a required feature in these linguistic contexts (Appel & Wood, 2016). Lexical chunks are typically defined as "groups of multiword units of language which are stored in long-term memory as if they were single lexical words" (Hou et al., 2016: 2). Lewis (1997) identified four main types of lexical chunks: words and poly words (e.g., *by the way*), collocations (e.g., *heavy rain*), institutionalized utterance frames (e.g., *We'll see*), and heads (e.g., *It is generally acknowledged that…*). Since UB approaches do not establish strict boundaries between grammar and lexicon subsystems (Römer, 2009), lexical chunks are seen as "conventionalized form–meaning mappings, the result of repeated use of certain linguistic units and specificity" (Hou et al., 2016: 3). From this perspective, then, lexical chunks are word sequences that can contain variable slots to be filled and used to express a concept in a conventionalized (i.e., formulaic) manner (Smiskova-Gustafsson, 2013). In the LA, learners need to be exposed to a significant stock of lexical chunks, which can be combined to produce "continuous, coherent text" (Lewis, 1993:7). If learners can remember lexical chunks after being exposed to considerable amounts of language (Durrant & Schmitt, 2010), then using prefabricated language can become a valuable tool in the learning process.

Advocates for the LA argue that classes focusing on learning as a product should be left behind, as teachers need to go beyond delivering content and completing textbook activities that highlight accuracy. Instead, this approach requires learners to engage in the observation of language features, to hypothesize on how such features can be used to communicate, and to experiment with language to communicate successfully. As learners experiment with language, it is likely that they will make grammatical mistakes in their production. These errors are regarded as an intrinsic feature of successful learning processes, as they evidence language development. Accordingly, Lewis (1993) states that learning is "essentially organic and involves connecting new materials to what is already known" (p. 56). Hence, the LA exposes learners to chunks, collocations and idiomatic expressions so that such structures can be internalized and subsequently accessed when needed. Furthermore, Lewis (1993) posited that language is mainly made of "grammaticalized lexis, not lexicalized grammar" (p. vi) and that "grammar as a structure is subordinate to lexis" (p. vii). These statements emphasize the importance of lexical phrases and their predominance over grammatical structures when communication is sought. Lewis argued for the need to remove grammar from

the central position it had previously held in traditional language learning approaches, and replace it with the use of functions (Lewis, 1993). For example, this would imply a shift from teaching the past simple to asking, "*What did you do yesterday?*". The lexical syllabus on which these practices must be based includes meaning-making activities that highlight the importance of using the language to acquire it, rather than the mere display of linguistic knowledge (Willis, 1990).

Another essential principle of the LA is the idea that language teaching must be focused on raising learners' awareness towards language chunking and how this impacts comprehension (Lewis, 1993). This implies that learners are expected to experience language and notice its features in context before they can acquire and use those features to communicate. When learning is seen as a process of language experimentation, the focus is not placed on how grammatically accurate an utterance is but on how the message is decoded by the receiver to establish successful communication. Although grammar explanations have been left aside in the implementation of the LA, learners may still notice syntactic patterns and rules as they experiment with language. These patterns may result in grammar formulations created by learners (i.e., grammatical rules such as third person *-s*) that must be valued by teachers even when they represent incorrect assumptions (Lewis, 1993).

1.1. Collocations

The role that frequent and natural language concordances – i.e., collocations – play in language perception and production is central in the LA approach. Lewis (2000) referred to collocations as "the way in which words co-occur in natural text in statistically significant ways" (p. 132). Shin & Nation (2008) defined them as "a group of two or more words that occur frequently together, and ... is not restricted to two- or three-word sequences" (p. 4). They also identified two parts in a collocation: "a pivot word which is the focal word in the collocation and its collocate(s), the word or words accompanying the pivot word" (p. 4). For example, in the sequences *electricity bill*, *gas bill* and *fuel bill*, the word *bill* is the pivot word, which is collocated by *electricity*, *gas* and *fuel*, respectively.

Collocations are a fundamental component of the LA. They play an important role in developing learners' oral proficiency and vocabulary strategies. One of the most important reasons for teaching collocations in language courses is that these structures can help develop fluency and native-like selection of language structures (Shin & Nation, 2008). This is a goal that, according to Siyanova & Schmitt (2008), "most advanced learners strive for, in their pursuit of second language proficiency, but few achieve" (p. 431). Benefits to using collocations have been reported by

Kashahara (2011), who found that EFL learners tend to remember more words when they appear next to other words because they facilitate the activation of already stored structures. However, teaching collocations may be challenging. For example, teachers must form noticing habits in their learners to raise awareness and help them identify word patterns in texts. Collocations must be carefully selected by teachers and must be appropriate and relevant to a particular proficiency level. Furthermore, learners must be encouraged to notice the collocations and guess the meaning of words and multi-word items by means of contextual features. These prefabricated structures are very diverse and can portray specific functions. Therefore, if they are not appropriately selected, they can fail to convey meaning and hinder language development. Boers & Lindstromberg (2009) suggest that in order to maximize the teaching of collocations, the proficiency level of the students and the aim of the lesson should be considered, as well as the usefulness and the frequency of the chunks.

1.2. Implementing the LA in the language classroom

As has been stated, the purpose of the LA is to expose learners to authentic language so as to increase their awareness and internalization of lexical chunks. As this type of input is crucial to the pedagogical process, the way in which teachers select and implement classroom materials plays a significant role in the approach. With respect to vocabulary sources, Vasiljevic (2014) highlights the importance of using dictionaries to learn about word sequence meanings and usages, along with training learners to use these tools regularly. Vocabulary learning materials should help learners in two important areas. First, materials must present learners with frequent practice of appropriate vocabulary in natural contexts, according to their language needs. Second, learners should adopt vocabulary learning techniques and strategies that can be used outside the classroom and in real language settings (McCarten, 2007). As Lewis (1993) argued, instead of using material specifically tailored for learners, there are "real materials" (p. 186) such as songs, TV programmes, and videos, that teachers can use to introduce lexical chunks. In the LA, the main role of teachers in this respect is to provide students with these materials to increase their awareness of how prefabricated structures are frequently used. Although non-native writers tend to overuse high-frequency collocations and underuse low-frequency ones (Durrant & Schmitt, 2009), elementary level learners can benefit from being exposed and using the former structures within an LA methodology. Furthermore, if long-term acquisition is facilitated by materials that are only partially understood (Lewis, 1993), it follows that real, authentic materials fit well in the LA. Nonetheless, Lewis (1993) warns that supplementary materials need to be prepared to support learning with authentic sources.

Lewis put forward a number of suggestions on how to tackle LA principles in the classroom. The LA requires that most of the classroom time be devoted to two important aspects in the approach: exposing learners to all types of multi-word items and raising their awareness of how language is used. In line with UB approaches, focusing on these aspects does not mean disregarding form, grammar, or creative use of language (Lewis, 1997). Teachers are allowed to address learners' questions about grammar patterns, but real-language materials have priority over descriptions on how language works (Lewis, 1993). In order for learners to acquire long-term grammatical accuracy, real-language awareness is heightened to make them notice collocational patterns instead of having teachers explain the underlying syntactic patterns. Authentic classroom language is vital in early lessons; common classroom-related chunks must be frequently used by the teacher so that leaners can familiarize with them and internalize them for later use. Expressions such as "*Sorry, I don't know*" and "*Sorry, I can't remember*", and sentence heads such as "*Did you...*" or "*Have you ever...*" are suggested in the approach. They must be taught as a whole and with no analysis at the word level. At early stages, teachers must emphasize language comprehension rather than production (Lewis, 1997). That is, teachers should not expect learners to use the prefabricated structures from the outset but should instead facilitate their repeated processing and understanding each time learners are exposed to them. As Lewis (1993) stated, "a short piece of real material, listened to several times is much more likely to be effective than a longer piece listened to once" (p. 186).

Lackman (2011) argues that words should not be taught without context in the LA. If lexical items are taught in isolation, it is less likely for those words to become part of the internalized chunk storage that teachers should seek to build. Lackman goes on to argue that idioms should not take up a large part of the lessons, as they are not commonly used by English speakers in real-life situations. In addition, encouraging a focus on collocation by means of dictionaries can lead to a successful implementation of the approach. Finally, Lackman suggests activities that can facilitate the implementation of the LA in the language classroom. Among these, the "*Find someone who...*" activity encourages students to find people who meet the requirements given in certain statements. For example, the teacher can ask students to look at a poster with famous people and ask them: "*Find someone who has written books*". Learners should then identify a particular person based on the descriptions. Another useful activity is "Giving clues", which requires the teacher to provide features of a lexical chunk so that learners can look for the structure that fits the description in the materials. For example, if the teacher asks for "a collocation made by a verb plus a noun", learners must underline "have dinner" in the printed material. Finally, "slot-fillers" can be used so that learners become aware of specific structures in multi-word items. Slot-filler prompts such as "*What did you* ____" can be completed with

many verb forms and adverbs: "*What did you do yesterday?*", "*What did you eat last night?*" (Lackman, 2011).

1.3. Previous studies on the Lexical Approach

Several studies have been carried out to investigate the effectiveness of the LA in the second language classroom. These studies have assessed the impact of exposing language learners to collocations and chunks on their written and oral production. The evidence provided by the literature comes from diverse contexts and is based on collocation-oriented perspectives applied in such settings. Research on the impact of the LA with a focus on formulaic sequences on college students' oral proficiency was conducted by Boers et al. (2006). The authors refer to formulaic sequences as "words or word string which appear to be processed without recourse to their lowest level of composition" (Wray, 2002: 4), and include other categories such as collocations, phrasal verbs, and idioms. Their 22-hour study consisted of an experimental group receiving exposure to real language and substantial amounts of authentic materials to enhance receptive skills (listening and reading) by means of phrase-noticing. The control group was taught in traditional lessons that differentiated grammar from vocabulary activities. Learners' performance in both groups was recorded and rated by judges. Results revealed that higher scores in perceived oral proficiency were obtained by students who were exposed to the LA. Boers et al. (2006) concluded that formulaic sequence instruction provides learners with sufficient vocabulary storage to allow them to store these multi-word items and retrieve them for immediate use. Similar results were reported by Stengers et al. (2011), who found a positive correlation between the frequency of formulaic sequence use by 60 Modern Languages students in Belgium and their level of oral proficiency. Wood (2006) conducted research on 11 English as a second language (ESL) learners to investigate the functions of formulaic sequences in speech fluency. He found that using these structures helped learners increase the range and use of functions when narrating stories. Similar findings were reported by Jones et al. (2015), who analyzed the chunks found in the spoken language of higher-intermediate level students. The oral production of 32 participants was assessed in terms of pronunciation, grammatical aspects, lexis, speech management, and ability to interact with language. Findings showed that the most frequent chunks used by learners at this proficiency level were often comparable to the chunks produced by native speakers, according to corpus data. Learners also used multi-functional chunks (e.g., *I think*) to convey a variety of meanings. The authors conclude that lexical chunks are more frequent at higher oral proficiency levels, and that teachers should focus on multi-functional chunks that can cover a wider range of meanings.

Balcı & Çakır (2012) reported positive effects when teaching vocabulary using collocations in a Turkish EFL classroom. The experimental group in their study was taught vocabulary by means of collocations, whereas the control group was taught with a traditional approach to vocabulary learning (i.e., by means of synonyms, antonyms, and definitions). Results suggested that vocabulary retention was higher in the group that was exposed to collocations. It must be noted that although learners usually create language patterns by means of collocations, these are not always accurate. According to Kuo (2009), learners make errors when selecting synonyms from the dictionary. When these synonyms do not collocate with the following word in a multi-word item, the resulting meaning becomes difficult or impossible to understand. These results are in line with Forteza et al. (2009), who found that learners tend to use synonyms to form incorrect collocations, which can hinder comprehension. Indeed, dictionary use may be insufficient if learners are only required to translate isolated words with no processing of the entire collocation. Focusing on L2 writing skills, Tang (2012) divided 85 second-year college students into an experimental (LA) group and a control group to gauge the impact of the LA on writing skills. Results revealed that LA is more effective in improving writing quality (measured by rating scores), as learners gather and internalize features from language and use them to produce written output. Thus, LA instruction allowed these learners to notice the existing gaps between the way they communicate and how native speakers typically perform. Tang concludes that exposing learners to prefabricated lexical sequences to communicate can reduce the number of incorrect sentences in written production, which can in turn increase accuracy and authenticity in the output.

Learners' perceptions in relation to the importance of teaching and learning collocations have also been reported in the literature. In this respect, Wu (2015) investigated the effect that teaching collocations has on the perceptions of elementary-level EFL students. Overall, findings showed that most participants reported positive attitudes toward the explicit teaching of collocations in the classroom. Learners stated that they felt more motivated to acquire new language structures, and that they acknowledged their own lack of interest toward words that are frequently accompanied by other words. In addition, they stated that it was confusing for them to understand why only certain words were followed or preceded by others (e.g., "powerful" preceding "engine" instead of "strong", even though both words are synonyms). They were aware of the instructor's efforts to encourage practice with the structures and to lower anxiety levels, which highlights the role that teachers play in the implementation of an approach that focuses on collocations. On the other hand, there were students who argued that while the concordance tools provided by the teacher throughout the treatment (e.g., collocation dictionaries and language databases) were helpful, they still found it problematic to use them because they were presented entirely in the L2.

Indeed, no access to L1 materials can confuse and frustrate learners at early stages of language development, and teachers should prepare supplementary materials to compensate for this. Overall, the reviewed studies underscore the value in teaching prefabricated sequences. Research findings suggest that this type of instruction should be implemented by language teachers to support learning and that researchers should continue to evaluate its impact on proficiency.

2. Research Methodology

The present study aims at assessing the impact of the Lexical Approach (LA) on adult EFL learners' perceived oral proficiency. The research questions are as follows:

1 What is the impact of the LA on the perceived oral proficiency of adult EFL learners?

2 What is the impact of the LA on the number of formulaic sequences used by these learners?

3 What is the relationship between perceived oral proficiency and the number of formulaic sequences used by these learners?

Research question 1 sought to explore how adult EFL learners use formulaic sequences in instructed communicative settings as part of the LA and to characterize the impact of teaching such sequences on their perceived oral proficiency (Boers et al., 2006; Millar, 2010). Research question 2 aimed at finding whether the LA would increase the number of lexical chunks produced by learners exposed to the LA. Finally, research question 3 sought to establish a relationship between the frequency of formulaic sequences used by these learners and their perceived oral proficiency (Stengers et al., 2011). The three research questions aimed at gauging the pedagogical strength of the lexical approach and contributing to the empirical evidence for the efficacy of UB approaches in EFL (Tyler et al., 2018).

The present study adopted a quasi-experimental design that included an intervention of 38 teaching hours over a four-month span. An experimental (LA) group and a control group were set up for the study, and a teacher in each group was instructed to adopt a particular teaching approach. Following Boers et al. (2006), the LA group was taught by means of formulaic expressions including lexical items, chunks, and collocations through real-language materials. The control group was exposed to regular lessons that took the same amount of teaching hours and were based on grammatical explanations, grammatical feedback, self-explanatory grammatical

handouts, worksheet completion, and general textbook activities. Collected data was analyzed to assess the impact of the LA on perceived oral proficiency, on the number of formulaic sequences utilized by learners, and on the relationship between the number of formulaic sequences used by learners and their oral proficiency.

2.1. Participants

The participants were 38 students (25 males and 13 females) in two intact adult EFL classes at a professional institute. Each class consisted of 19 EFL students who were enrolled in a compulsory elementary EFL course as part of their computer engineering programme. Their ages ranged from 18 to 25 years old, and their CEFR (Common European Framework of References for Languages) oral proficiency level was elementary English (A2), as measured by a final oral exam completed at the end of their previous course. The EFL courses in the institute sought to enhance students' communicative skills to further career development.

2.2. Treatment

The experimental group (LA) was taught using LA principles based on co-occurring lexical units, whereas the control group received regular lessons. Both groups were taught the same contents, as stated in the course syllabus. In the LA group, Lackman's (2011) suggestions were followed. Authentic materials such as published songs, newspapers and TV show videos were used, and learners were not exposed to worksheets with grammatical explanations. As can be seen in the sample activities (Appendix), classroom tasks such as "*Find somebody who...*" were adapted and implemented in the LA group. Students were given songs and asked to look for a sentence that would fit the linguistic clue given by the teacher. Songs are "a rich source of lexical items, particularly commonly used semi-fixed expressions" (Lackman, 2011: 13). Slot-fillers were also implemented in the LA group. For example, group conversations about past experiences were held by using ready-made slot-fillers on the whiteboard. Structures such as "*Did you _____ last weekend?*" and "*Where did you _____ yesterday?*" were used to provide learners with sufficient language to ask questions to their classmates. Learners answered these questions by means of slot-fillers such as "*Yesterday, I _____*". The use of dictionaries was also encouraged as learners completed the lexical slots. As indicated by Lewis (1993), learners were asked to keep a vocabulary notebook and jot down useful phrases and words. Learners were advised to divide the notebook into sections (e.g., "giving directions" and "giving opinions") so that they could find new expressions and collocations more easily. They were also given a list with the vocabulary items reviewed in the lessons and were told to include them

in their notebooks. Learners in the LA group were encouraged to notice language features in every piece of authentic language to which they were exposed and to look up those lexical chunks in a dictionary. Drawing from Hsu & Chiu (2008), direct instruction of lexical collocations was implemented in the LA group, as it was found to be beneficial when attempting to teach vocabulary.

Although a coursebook was used to deliver the lessons in both the treatment and control groups, the approach taken with it was tailored to each condition. In the LA group, the activities chosen from the coursebook were carefully tailored to follow LA principles. Thus, coursebook sections that were directly related to grammar explanations were taken out, as syntactic patterns are expected to be noticed, but not explained to learners in the LA. In addition, "unauthentic" dialogues (Aston, 2009) were replaced with real-world materials. In the control group, learners had access to coursebooks, handouts and e-classes with summarized versions of every unit, L1 explanations, and content activities. The control group completed all the activities in the coursebook, which included sections such as the "grammar bank" – where grammatical structures are explicitly provided to students – and the "vocabulary bank" – where lists of isolated words reviewed in the lesson are presented. Furthermore, both groups differed in terms of the increased awareness toward lexical chunks in the activities completed by the LA (experimental) group. Teacher feedback in both groups was focused on providing meanings to single or multi-word structures, but participants in the LA group were asked to contextualize such meanings as part of lexical chunks. As reported by the teachers, the LA group did not spend more time on oral practice than the control group, as both teachers included between 15 and 20 minutes of conversational practice per lesson. However, the oral practice in the LA group was based on authentic materials (i.e., songs and videos), while the control group focused on practicing grammar structures. Table 1 below summarizes the type of materials and approaches implemented in both groups.

Table 1. Approaches and type of materials in the LA group and the control group

Group	LA Group	Control group
Materials (both groups were exposed to the same course content)	Coursebook (vocabulary activities with the LA approach), notebook for useful phrases and sentences.	Coursebook (vocabulary and grammar activities) and handouts.
Approach to grammar	No explicit focus on grammar.	Explicit treatment of grammar explanations and grammatical patterns.
Activities	-Authentic materials (i.e. songs, newspapers, videos). -Example of activity with a song: Students must find a sentence where the singer asks where something is. -Slot fillers (see Appendix). -Repeated exposure to lexical chunks. -Suggested vocabulary. -Oral practice with selected lexical chunks.	-Non-authentic material based on the coursebook: A: *Hello.* B: *Hello.* A: *What are you doing? Are you listening to music?* B: *No, I am not.* A: *Are you cooking?* B: *No, I am not.* -Grammar bank. -Grammar worksheets. -Coursebook vocabulary activities. -Oral practice with grammar points.

2.3. Procedures

2.3.1. Oral tests

As a pre-test measure to control for oral proficiency, we considered a final oral exam (oral assessment) that the participants completed at the end of the previous course level. This oral assessment required learners to complete a 10-minute interview with the teacher about coursebook topics. The rubric assessed range of oral expression, accuracy, and fluency, and the scores were used by the teacher as a final course grade for the learners. A Mann-Whitney U test was run to confirm that the oral proficiency of both groups was similar at the start of the treatment. The scores were not found to be significantly different ($p > .05$).

Before collecting post-test oral proficiency data, we secured consent from participants in the study. We provided them with an information sheet where the study was described, and anonymity and confidentiality issues were discussed. They were also told that they could withdraw from the study at any point in the data collection procedures. Once the consent forms were collected, participants completed the 38 weeks of treatment and then were asked to sit an oral test interview carried out by their teacher. A second teacher was present in the room to record the oral test. Participants were assessed individually in 10-minute sessions, which were divided into two tasks. In Task 1, learners had to role-play a situation in which they had to provide directions. Learners had to use a map to indicate how to get from one point to several other points. The teacher gave learners one minute to look at the map in order for them to get acquainted with the names of the streets and locations before giving directions. Task 2 required learners to answer questions using time expressions included in the vocabulary that was taught in the course, and to talk about a topic of their choice.

2.3.2. Judges and rubrics

The oral performance data was assessed by three judges. The judges were two native speakers of English who worked as EFL teachers in language institutes and an experienced non-native EFL teacher. The two native speakers of English (Judge A and Judge B) were asked to assess the oral proficiency of participants using two different CEFR rubrics (Council of Europe, 2001). The first rubric focused on qualitative aspects of spoken language and was used by the judges to assess learners' performance considering range of expression, accuracy, and fluency. This measure has been used in other studies addressing perceived learners' oral proficiency (Boers et al., 2006; Stengers et al., 2011). The second rubric measured overall spoken production, which provided a more holistic assessment of the recordings. Each of the components in the rubrics had the same weight in the ratings provided by the judges. Using the mean scores for the components in both rubrics, Judge A and Judge B had to assign a final numeric score that represented the CEFR level attained by a participant (from 1-3 for A1, to 15-18 for C2). Finally, the experienced non-native EFL teacher (Judge Y) was given a checklist sheet and was asked to list and count the formulaic sequence tokens identified in the participants' recordings. Following Boers et al. (2006), Judge Y was asked to count the formulaic sequences without making any judgements on their complexity, because doing so would bring about qualitative aspects that are difficult to assess objectively. She was familiar with the concept of formulaic language and with the literature on lexical chunks, and she did not know whether the participants belonged to the experimental or the control group.

Two reliability measures were adopted regarding the scores provided by the judges. With respect to the oral proficiency scores provided by Judge A and Judge B, intraclass correlation coefficients were calculated. They were found to be over .96 across all components, which showed that the judges were highly consistent in the scores they provided. Hence, an average score between Judge A and B was computed and considered for analysis. Likewise, an overall proficiency score for Task 1 and Task 2 was computed and taken as a measure of oral proficiency. In order to provide a measure of intra-rater reliability for Judge Y, three weeks after her first formulaic sequence count, she was asked to go through the process again. The intraclass correlation coefficient for the formulaic sequence counts (.938) suggested excellent reliability.

3. Results

3.1. Descriptive statistics for oral proficiency scores

Table 2 displays descriptive statistics for oral proficiency scores (range of expression, accuracy, fluency, and overall proficiency) in both groups, as perceived by both judges (Judge A and Judge B) as an average for both tasks (Task 1 and Task 2). The means in the control group for all measures were more than doubled in the experimental group. As data did not display normal distribution, the median (Mdn) and the interquartile range (IQR) were used as measures of central tendency and dispersion. Across measures, the medians in the LA group were considerably higher than in the control group. The medians for specific measures in the control group ranged from 3.75 to 4.5, while the same medians in the LA group ranged between 9.75 and 11.25.

Table 2: Descriptive statistics for oral proficiency scores across measures and groups

Group	Measures	N	Minimum	Maximum	Mean	Mdn	IQR
LA	TASKS_RANGE	19	5.5	17.75	10.82	11.25	5.75
	TASKS_ACC	19	5.25	15.75	10.07	9.75	4.75
	TASKS_FLU	19	5	17.75	10.53	10.75	5.75
	TASKS_OVERALL	19	4.50	16.75	10.12	10.00	5.5
Control	TASKS_RANGE	19	1	12.25	5.16	4.5	5.25
	TASKS_ACC	19	1.25	12.75	4.9	3.75	4.5
	TASKS_FLU	19	1.25	11.25	4.84	4.25	3.5
	TASKS_OVERALL	19	1.25	11	4.93	5	4

RANGE: Range of expression. ACC: Accuracy. FLU: Fluency. OVERALL: Overall oral proficiency.

3.2. Mann-Whitney U tests for overall judge scores

To test whether oral proficiency scores were statistically different in both groups across oral proficiency measures (Tasks 1 and 2) provided by the average proficiency scores (Judges A and B), Mann-Whitney U tests were run (Table 3). Results revealed that the differences in perceived oral proficiency between both groups were statistically significant in specific and overall measures ($p < .05$).

Table 3: Mann Whitney U tests for average oral proficiency scores across measures

Variable	U	*p*	Mean rank LA group	Mean rank control group
Range	46.5	.000	26.55	12.45
Accuracy	45.5	.000	26.61	12.39
Fluency	41	.000	26.84	12.16
Overall	53.5	.000	26.18	12.82

3.3. Descriptive statistics for the formulaic sequence count

Descriptive statistics for the formulaic sequences (FSs) identified by Judge Y in both groups were computed considering the two tasks as separate measures (Table 4 below) because they revealed higher means and medians for the FSs in both groups in Task 1 (giving directions) in comparison to Task 2 (answering questions and self-selected topic). Thus, both groups found it more difficult to produce FSs in Task 2, regardless of group condition.

Table 4: Descriptive Statistics for formulaic sequence count

Group	TASK	N	Minimum	Maximum	Mean	Mdn	IQR
LA	1	19	3	7	5.26	5.00	2
	2	19	1	3	2.16	2.00	1
Control	1	19	3	6	4.32	5.00	2
	2	19	1	4	1.89	2.00	1

3.4. Mann Whitney U tests for Judge Y's FS counts

We further explored the difference in medians for Task 1 and Task 2 reported above by means of Mann-Whitney U tests. As can be seen in Table 5, statistical analysis yielded different results depending on the task being assessed. Results in Task 1 display a significant difference in the FS counts between both groups (p = .019), with a moderate effect size (r= 0.38). This difference between groups did not reach significance in Task 2 (p < .05).

Table 5. Mann Whitney U tests for Judge A's FS counts in both groups

Judge	Task	U	*p*	Mean rank LA group	Mean rank control group
Y	1	103	.019	23.58	15.42
Y	2	142	.220	21.55	17.45

3.5. Correlations between overall oral proficiency scores and formulaic sequence counts

This section presents the correlations performed between the overall oral proficiency scores (Task 1 and Task 2) of participants as perceived by Judges A and B (average score) and the mean number of FSs counted by Judge Y in those tasks. These correlations sought to assess the extent to which the improvement in terms of oral proficiency is related to the number of FSs produced by the participants. For this purpose, Spearman rank tests were run, as data was found to be non-parametric. Table 6 shows that moderate significant positive correlations between the number of FSs counted and oral proficiency scores were found. These correlations ranged from .478 (range of expression and FS counts) to .497 (fluency and FS counts).

Table 6. Spearman's rho correlations between average oral proficiency score measures and FS counts

	RANGE	ACC	FLU	OVERALL
FSC	,478*	,491*	,497*	,478*

FSC: Formulaic sequence count. RANGE: Range of expression. ACC: Accuracy. FLU: Fluency. OVERALL: Overall oral proficiency.

*Correlation is significant at the 0.01 level (2-tailed).

4. Discussion

4.1. What is the impact of the LA on the perceived oral proficiency of adult EFL learners?

The first research question aimed to assess the impact of the LA on learners' perceived oral proficiency. Results revealed that participants' oral proficiency scores were significantly higher in the LA group, which suggests that the implementation of the LA principles in the treatment condition had a positive effect on learners' oral proficiency scores in terms of range of expression, fluency, and accuracy. These results are in line with Boers et al. (2006) and Stengers et al. (2011), who reported higher oral proficiency scores in those components when learners were taught using formulaic expressions. The number of lexical chunks delivered through authentic language in the LA group may have raised learners' awareness toward those structures and facilitated their use when producing more accurate and fluent language (Boers & Lindstromberg, 2009; Lewis, 1997). These results also contribute to the literature reporting the positive effects of teaching formulaic sequences and collocations on the development of fluency and a native-like selection of structures (Shin & Nation, 2008; Wood, 2006, 2010). In particular, the perceived oral proficiency scores given by Judges A and B in the present study may have been influenced by the native-like selection of lexical chunks done by participants in the LA group. Results are also in line with Serrano et al. (2015), who found that lower and intermediate level learners benefit the most from formulaic sequence instruction.

4.2. What is the impact of the LA on the number of formulaic sequences used by these learners?

The second research question sought to measure the impact of the LA on the number of formulaic sequences used by these participants in the tasks. Significant differences were found in Task 1, where the LA group outperformed the control group on the number of FSs used to give directions. Although the number of formulaic sequences produced in Task 2 was higher in the LA group, this difference was not significant. These results can be discussed in terms of the impact of the type of task on participants' performance. Boers et al. (2006) reported that their participants performed differently according to the task they were facing (conversation about a magazine article and spontaneous conversation about a familiar topic). In line with the findings in the present study, Boers et al. found that participants used more formulaic expressions in the LA group when they were engaged in a conversation about a magazine article. Their participants may have been able to notice lexical chunks and use them

when discussing the magazine article because they were given time to read it. This also happened in Task 1 (Giving directions) in the present study, as participants were given some time to look at the map and get acquainted with the streets and landmarks before starting the role-play task. While Task 1 addressed role-play practice with more specific and predetermined sequences for learners to use, Task 2 was less specific in this respect and required learners to speak about a topic they had just chosen and without preparation. Thus, participants may have produced fewer formulaic sequences in Task 2 because in Task 1 they had time to process specific lexical chunks and their language did not require complex lexical or syntactic modifications to be successfully used (*turn left, turn right, go straight ahead*). This seems to be in line with findings suggesting that producing formulaic sequences has a processing advantage over language that is generated through speakers' creativity (Conklin & Schmitt, 2008). Furthermore, awareness raising is one of the most important features in the LA (Lewis, 1993), so it can be argued that the non-significant results found in Task 2 may be due to the limited number of hours devoted to this kind of task. As Rafievan (2018) argues, "classroom time is clearly too limited to explicitly teach more than a fraction of the vast number of formulaic sequences in language" (p. 2). Indeed, the amount of time devoted to the LA and its features must aim at long-term objectives to achieve successful implementation. The treatment in the present study included 38 hours of LA practice, while Boers et al.'s intervention (2006) consisted of 22 hours. Still, a more comprehensive approach to formulaic sequences might provide more measurable differences with respect to spontaneous speaking tasks.

4.3. What is the relationship between perceived oral proficiency and the number of formulaic sequences used by these learners?

The third research question aimed at assessing the relationship between the participants' perceived overall oral proficiency (rated by Judges A and B) and the number of formulaic sequences they used (counted by Judge Y). Findings revealed moderate positive correlations across all components of oral proficiency; that is, the higher the number of formulaic sequences used in the tasks, the higher the oral proficiency scores given by the native-speaker judges. Overall, the perceived proficiency level of participants was found to be related to the number of formulaic sequences produced by them, results that are in line with Stengers et al. (2011). The findings of the present study are also in line with Wood (2006), who found that participants who performed better in fluency measures tended to use more formulaic sequences in their speech. These findings indicate that mastering ready-made structures when learning a new language will likely increase perceived oral proficiency.

5. Conclusion

In the present study, three main findings can be reported in relation to the LA and its effects on oral proficiency and FSs. First, the implementation of the LA based on frequent exposure to form-meaning associations advocated by UB approaches (Tyler et al., 2018) was found to have a positive impact on the development of learners' oral proficiency as perceived by native-speaker judges. Thus, language development can be, to some extent, promoted by repeated exposure to authentic language and awareness raising toward lexical bundles. Second, LA principles can influence the development of a formulaic sequence repertoire in learners, particularly in the production of specific lexical chunks. Even though the LA group outperformed the control group in terms of formulaic sequence counts in both tasks (giving directions and questions and answers/ talk about), this difference only reached significance in Task 1, where participants were asked to give directions using a map. This finding highlights the importance of allocating enough practice time to the development of a multi-word vocabulary repertoire in learners that can be retrieved for spontaneous talk. Finally, the moderate correlations found suggest that there is a relationship between the number of formulaic sequences used and perceived oral proficiency. In other words, using more formulaic sequences will likely increase the perceived oral proficiency score of a learner. This analysis was done by focusing on the number of formulaic sequences produced (see Boers et al., 2006) rather than on their complexity (i.e., length of the formulaic sequence), due to the limited oral proficiency in these learners. For this reason, a straightforward measure of frequency was selected over a total number of words/words in FS ratio (cf. Hou et al., 2016). Nonetheless, the inclusion of complexity measures in future studies could yield additional data characterizing the use of formulaic sequences in higher proficiency level learners. In addition, although written production was not assessed in the present study, it is likely that repeated exposure to lexical chunks can have a positive impact on the written output of learners when compared to translating isolated words from a dictionary (Forteza et al., 2009; Kuo, 2009). However, studies should include linguistic measures of written output (i.e., syntactic complexity) that go beyond holistic rubrics (see Cancino & Panes, 2021) to assess specific aspects of proficiency and confirm the impact of the LA on written production.

5.1. Implications for pedagogy

The pedagogical implications of the findings in the present study can be addressed in terms of the effectiveness of a teaching approach that is based on LA principles. The treatment group was exposed to 38 hours of LA instruction, and results provide evidence highlighting the benefits of teaching language items as wholes, rather than spending most of the classroom time on grammar explanations (as was the case in the

control group). However, it would seem that fully applying LA principles in contexts that focus on the short-term achievement of grammatical accuracy may become problematic. That is, a primary focus on grammar accuracy will likely disregard the emphasis on long-term internalization of lexical chunks by means of authentic materials proposed by the LA approach. Some level of accuracy is necessary to achieve comprehensibility in the LA, but this is neither essential nor expected at early stages of second language development. This stance is in accordance with the more implicit approach to grammar advocated by usage-based pedagogies (Tyler et al., 2018). Moreover, the limited potential exposure to multi-word form-meaning structures that EFL learners in the selected context receive (i.e., two 90-minute lessons a week) hinders the opportunities for noticing and processing lexical chunks. Therefore, EFL teachers need to maximize their learners' practice with these structures so that the time between each exposure is reduced and internalization can take place (Serrano et al., 2015).

The present study contributes to the evidence showing that instruction focusing on raising awareness toward co-occurring lexical units can increase oral proficiency over time. This awareness towards lexical chunks can be boosted by exposing learners to authentic language materials, rather than to the sheltered materials typically found in textbooks. Furthermore, natural rhythm in speech is more likely to be found in the real-life language presented in teaching methodologies based on the LA, which can increase awareness toward authentic speech patterns. As the LA requires learners to notice patterns and experiment with them without engaging in short-term grammar assessment (Lewis, 1993), long-term goals for the teaching of formulaic expressions need to be set if the approach is to yield oral proficiency development.

6. References

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels. *Language Assessment Quarterly, 13*, 55-71.

Aston, G. (2009). *Corpora and language learners*. Amsterdam: Benjamins.

Balcı, Ö., & Çakır, A. (2012). Teaching vocabulary through collocations in EFL Classes: The case of Turkey. *International Journal of Research Studies in Language Learning, 1*(1), 21-32.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245-261.

Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Hampshire, UK: Springer.

Cancino, M., & Panes, J. (2021). The impact of Google Translate on L2 writing quality measures: Evidence from Chilean EFL high school learners. *System*. Advance online publication.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*, 72-89.

Council of Europe. (2001). *Common European framework of references for languages. Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, *47*, 157-177.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure.*Second Language Research*, *26*, 163-188.

Ellis, N. C. (2013). Construction grammar and second language acquisition. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 365-378). New York: Oxford University Press.

Forteza, R., Prahlad, S. R., Rubtsova, E., & Sabitov, O. (2009). Collocations in the vocabulary English teaching as a foreign language. *Acimed, 19*(6), 1-5.

Hou, J., Loerts, H., & Verspoor, M. H. (2016). Chunk use and development in advanced Chinese L2 learners of English. *Language Teaching Research, 22*(2), 148-168.

Hsu, J.-Y., & Chiu, C.-Y. (2008). Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *Asian EFL Journal, 10,* 181-204.

Jones, C., Waller, D., & Golebiewska P. (2015). Spoken lexical chunks used by successful learners at B2 level: Forms and functions. *CALR Linguistics Journal*, 1-13.

Kasahara, K. (2011). The effect of known-and-unknown word combinations on intentional vocabulary learning. *System*, *39*, 491-499.

Kuo, C. (2009). An analysis of the use of collocation by intermediate EFL college students in Taiwan. *ARECLS, 6*, 141-155.

Lackman, K. (2011). *Lexical Approach Activities*. Ken Lackman & Associates Educational Consultants. Retrieved from http://kenlackman.com/files/LexicalActivitiesBook102.pdf

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Hove: Language Teaching Publications.

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice. Language.* Hove: Language Teaching Publications.

Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach.* Hove: Language Teaching Publications.

Madlener, K. (2018). Do findings from artificial language learning generalize to second language classrooms? In A. Tyler, L. Ortega, M. Uno, & H.I. Park (Eds.). *Usage-Inspired L2 Instruction* (pp. 211–234). Amsterdam: John Benjamins Publishing Company.

McCarten, J. (2007). *Teaching vocabulary: Lessons from the corpus, lessons for the classroom*. New York: Cambridge University Press.

Millar, N. (2010). The processing of malformed formulaic language. *Applied Linguistics*, *32*, 129-148.

Rafieyan, V. (2018). Role of knowledge of formulaic sequences in language proficiency: Significance and ideal method of instruction. *Asian-Pacific Journal of Second and Foreign Language Education, 3*(1), 1-23.

Römer, U. (2009). The inseparability of lexis and grammar. *Annual Review of Cognitive Linguistics*, 7, 140–162.

Serrano, R., Stengers, H., & Housen, A. (2015). Acquisition of formulaic sequences in intensive and regular EFL programmes. *Language Teaching Research, 19*(1), 89–106.

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339-348.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review, 64*(3), 429-458.

Smiskova-Gustafsson, H. (2013). *Chunks in L2 Development: A usage based perspective*. Groningen: Grodil.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics*, *49*, 321-343.

Tang, J. (2012). An empirical study on the effectiveness of the lexical approach to improving writing in SLA. *Journal of Language Teaching & Research, 3*(3), 578-583.

Tyler, A. E., & Ortega, L. (2018). Usage-inspired L2 instruction. In A. Tyler, L. Ortega, M. Uno, & H.I. Park (Eds.). *Usage-Inspired L2 Instruction* (pp. 3-26). Amsterdam: John Benjamins Publishing Company.

Tyler, A. E., Ortega, L., Uno, M., & Park, H. I. (Eds.). (2018). *Usage-inspired L2 instruction: Researched pedagogy* (Vol. 49). Amsterdam: John Benjamins Publishing Company.

Vasiljevic, Z. (2014). Teaching collocations in a second language: Why, what and how. *ELTA Journal, 2*(2), 48-73.

Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. London: Harper Collins Publishers.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*(1), 13-33.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. New York: Continuum.

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

Wu, K. (2015). The effect of teaching collocations on the students' perceptions toward language learning. *Studies in English Language Teaching, 3*(3), 215-236.

Appendix.

Sample activities in the LA group

1) **Read the following clues. Find and underline…**
   a. A sentence to ask where a place is
   b. An expression to express that someone's opportunities started to decrease.
   c. A sentence to say that someone took a longer path.
   d. An action performed in a car.
   e. A sentence to describe the physical appearance of a woman.
   f. A sentence to express someone's worry about a future event.

2) **Can you create more clues? Write 3 more clues to describe sentences from the lyrics. Then, tell them to your classmate. Check if he/she can guess.**
   a. _____
   b. _____
   c. _____

3) **Complete the following sentences from the song. Try it without looking at the lyrics. Do you remember the words?**

   a. I _____ I _____ you in the battleship.

   b. When I _____ her if I _____ call her your name.

   c. She was close, and she _____ me very tightly.

   d. I'm beginning to think I _____ you all along.

   e. I _____ your scent on the seat belt.

4) **Could you remember them? Now, use your dictionary and find more words to fill the gaps. Do they have the same meaning? Did the meaning of the sentence change?**

   Example:

   a. *I <u>believed</u> I <u>watched</u> you in the battleship.*

   b. She _____ close, close enough to be your ghost.

   c. When I _____ her if I _____ call her your name.

   d. I _____ over for a closer look.

   e. She was close, and she _____ me very tightly.

   f. And I _____ my lift home.

   g. I'm beginning to think I _____ you all along.

   h. I _____ your scent on the seat belt.

5) **Find a song about past experiences and print the lyrics.**

   a. Create 3 clues to describe sentences in the lyrics.

      i. _____

      ii. _____

      iii. _____

   b. Tell the clues to your classmate while he/she is listening and reading the song. Can he/she guess them?

c. Find 3 sentences that describe past experiences and find synonyms for the verbs. Do not change the meaning of the sentence!

i. _____

ii. _____

iii. _____

Incidental Focus-on-Form Characteristics: Predicting Learner Uptake. Formulaic vs. Non-Formulaic Forms

Leila Gholami
Department of English
Arizona State University
lgholami@asu.edu

Abstract

A substantial number of studies have investigated the efficacy of incidental focus on form (FonF) measured through (successful) uptake rate in teacher-learner interactions in communicative contexts and have established a link between learners' (successful) uptake of linguistic forms and their second language learning. In this line of research, the analysis of uptake and FonF characteristics mediating learners' (successful) uptake has been limited to linguistic forms of grammar, vocabulary, pronunciation, and spelling (non-formulaic forms). However, formulaic sequences, including idioms, collocations, lexical bundles, and compounds (formulaic forms), have received scant attention in FonF research. This study examined incidental FonF characteristics that best predicted learners' (successful) uptake of formulaic forms and compared the findings with the variables that mediate the (successful) uptake of non-formulaic forms targeted in focus-on-form episodes (FFEs). To this end, 30 hours of audio-recorded teacher-learner interactions in primarily communicative activities from English as a foreign language classes were examined. The findings showed that learners' attention was drawn to non-formulaic forms more frequently than formulaic forms in FFEs. Nonetheless, learners produced (successful) uptake more often when formulaic forms were targeted in FFEs than non-formulaic forms. Logistic regression analyses showed that FonF characteristics that predicted learners' production of (successful) uptake were different for formulaic vs. non-formulaic forms targeted in FFEs.

Keywords: EFL; Focus on form; formulaic sequences; incidental FonF characteristics; successful uptake.

Resumen

Un número considerable de estudios ha investigado la eficacia del enfoque incidental en la forma (FonF) medida a través de la tasa de asimilación (exitosa) en las interacciones entre el profesor y el alumno en contextos comunicativos y han

establecido un vínculo entre la asimilación (exitosa) de formas lingüísticas por parte de los alumnos y su aprendizaje de una segunda lengua. En esta línea de investigación, el análisis de las características de asimilación y FonF que median en la asimilación (exitosa) de los alumnos se ha limitado a las formas lingüísticas de gramática, vocabulario, pronunciación y ortografía (formas no formulaicas). Sin embargo, las secuencias de fórmulas, incluidas las expresiones idiomáticas, las colocaciones, los conjuntos léxicos y los compuestos (formas de fórmulas), han recibido escasa atención en la investigación del FonF. Este estudio examinó las características incidentales del FonF que mejor predijeron la asimilación (exitosa) de las formas formulaicas por parte de los alumnos y comparó los resultados con las variables que median en la asimilación (exitosa) de las formas no formulaicas que son objeto de episodios de enfoque en la forma (FFE). Para ello, se examinaron 30 horas de interacciones profesor-alumno grabadas en audio, principalmente en actividades comunicativas de clases de inglés como lengua extranjera. Los resultados mostraron que la atención de los alumnos se dirigía a las formas no formulaicas con más frecuencia que a las formulaicas en los episodios de FonF. Sin embargo, los alumnos producían una captación (exitosa) más a menudo cuando las formas formulaicas eran el objetivo de los FFEs que cuando lo eran las formas no formulaicas. Los análisis de regresión logística mostraron que las características del FonF que predecían la producción de asimilación (exitosa) por parte de los alumnos eran diferentes para las formas formulaicas frente a las formas no formulaicas dirigidas a los episodios de enfoque en la forma (FFE).

Palabras clave: ILE; enfoque en la forma; secuencias formulaicas,;características de enfoque en la forma incidentales; captación exitosa.

1. Introduction

Incidental focus on form (FonF) refers to directing learners' attention to linguistic forms as they arise spontaneously in primarily communicative interactions (Long, 1996). Successful uptake is used in descriptive FonF studies as a metric for the potential efficacy of incidental FonF in facilitating second language (L2) learning (e.g., Ellis, Basturkmen, & Loewen, 2001; Li & Vuono, 2019; Loewen, 2004). Uptake is characterized as "learners' responses to the provision of feedback after either an erroneous utterance or a query about a linguistic item within the context of meaning-focused language activities" (Loewen, 2004: 153). Successful uptake refers to learners' target-like modification of their utterances following FonF, and learners' non-target-like modification of their original output is termed as unsuccessful uptake (Egi, 2010). Successful uptake is "*facilitative* of acquisition" (Ellis et al.: 287, emphasis in original) by "providing opportunities for learners to proceduralize target language knowledge already internalized in the declarative form" (Lyster, 1998: 191). According

to Schmidt's (1995) noticing hypothesis, linguistic forms that are noticed are learned more effectively. Learners' production of successful uptake following FonF has been found to be the evidence of their noticing target forms (e.g., Egi, 2010; Gurzynski-Weiss & Baralt, 2015).

A number of variables such as the type of FonF, the timing of FonF, type of corrective feedback, and other pedagogical choices in the provision of incidental FonF have been found to influence the rate of uptake and its quality (i.e., successful and unsuccessful uptake) (Ellis et al., 2001; Loewen, 2004). Previous FonF studies have examined FonF, uptake, and the characteristics that mediate learners' production of (successful) uptake in linguistic forms limited to grammar, vocabulary, pronunciation, and spelling (non-formulaic forms) (e.g., Ellis et al., 2001; Loewen, 2004). However, to date, no studies have examined FonF characteristics predicting learners' production of (successful) uptake in relation to formulaic sequences (FSs) including idioms, collocations, lexical bundles, and compounds (formulaic forms).

FSs are described as "phrases that are conventional pairings of forms with units of meaning in a speech community" (Buerki, 2016: 21). The significance of FSs in L2 development lies in the pervasiveness of FSs in natural language discourse, communicative functions of FSs, rapid processing of FSs, and the use of FSs as indicative of native-like proficiency (Wray, 2019). It is important to examine the variable predicting learners' (successful) uptake of FSs because of the key role of FSs in communicative language use (Wray, 2019) and the association between learners' successful uptake of linguistic forms and their L2 learning (Loewen, 2004). Therefore, this study examined incidental FonF characteristics that best predicted learners' (successful) uptake of formulaic forms and compared the findings with the variables that mediate the (successful) uptake of non-formulaic forms targeted in incidental FonF.

2. Literature review

2.1. Focus on form and uptake

Incidental FonF is divided into reactive and preemptive types (Ellis et al., 2001). In reactive FonF, learners are provided with corrective feedback in response to their non-target-like output (Loewen, 2004). In preemptive FonF, either the learner (student-initiated) or the teacher (teacher-initiated) takes the initiative to shift learners' attention to forms by raising a query or making a comment on a form regardless of an error occurrence (Loewen, 2004). The efficacy of incidental FonF in promoting L2 learning has been measured by uptake and successful uptake rate in descriptive studies (e.g., Ellis et al., 2001; Loewen, 2004; Lyster, 2001).

The significance of (successful) uptake in L2 learning largely derives from the role of noticing (Schmidt, 1995) and pushed output (Swain, 1995) assumed in L2 development. Swain (1995) argues that pushing learners to produce language encourage them to notice the gap in their interlanguage. Schmidt (1995) maintains that learners' noticing the mismatch between their interlanguage and the target language is conducive to restructuring interlanguage toward target forms. The learners' production of (successful) uptake following FonF is a form of pushed output (Egi, 2010) and the evidence of their noticing target forms (Gurzynski-Weiss & Baralt, 2015).

The incidence of uptake and its quality (i.e., successful and unsuccessful uptake) have been found to be contingent on different pedagogical choices in the provision of incidental FonF (e.g., Ellis et al., 2001; Loewen, 2004). Investigating form limited to vocabulary, grammar, and pronunciation, Ellis et al. (2001) found that (successful) uptake rate varied depending on the type of incidental FonF (i.e., reactive vs. preemptive FonF), the source of FonF (i.e., a problem with communication vs. a problem with the accuracy of the form), and complexity of FonF (i.e., whether attention to form involved several complex moves vs. simple moves). Loewen (2004) found that characteristics such as type of feedback (eliciting target forms from learners vs. providing them with target forms) and timing of FonF (immediate vs. delayed) influenced both the production of uptake and the successfulness of it. The investigation of FonF characteristics that mediate learners' production of (successful) uptake following FSs is a gap in FonF studies.

2.2. Formulaic sequences

Various types of FSs include collocations (*make money*), idioms (*make a killing*), proverbs (*let's make hay while the sun shines*), binomials (*bride and groom*), lexical bundles (*as a consequence*), compounds (*chain store*), and pragmatic formulas (*nice to meet you*) (Siyanova-Chanturia, 2019). FSs are pedagogically important for L2 development based on the following findings. The findings of corpus linguistics attest that FSs are prevalent in natural language use (Erman & Warren, 2000). Pragmatic formulas (*how do you do?*) play a key role in performing pragmatic and discourse functions (Kecskes, 2016). The use of FSs fosters fluency as FSs are processed holistically, obviating the need to generate language in a word-for-word fashion from scratch (Siyanova-Chanturia, 2019). FSs also maximize accuracy as FSs are fixed and long chunks allowing for few or no modifications (Wray, 2019).

However, there is a consensus that even advanced L2 learners' knowledge of formulaic language lags behind that of grammar and single-word vocabulary (Meunier, 2012; Sinclair, 1991; Wray, 2019). Meunier (2012) argues that despite the importance

of FSs in L2 development, they have not figured prominently in language pedagogy due to the traditional emphasis on vocabulary and grammar. One way to facilitate L2 learners' acquisition of FSs is to raise their attention to FSs through incidental FonF and to promote learners' noticing and (successful) uptake of target FSs. Gholami and Gholami (2018) investigated the occurrence of (successful) uptake in 36 hours of communicative interactions and found that learners tended to produce (successful) uptake more often following target formulaic forms than non-formulaic ones. Investigating linguistic form limited to grammar, pronunciation, and vocabulary, Ellis et al. (2001) and Loewen (2004) found FonF characteristics such as the timing of FonF (immediate vs. delayed), type of FonF (reactive vs. preemptive), type of corrective feedback (elicitation vs. provision of correct target form), etc., mediated the occurrence of (successful) uptake in incidental FonF. To the best of the researcher's knowledge, to date, no studies have examined FonF characteristics that mediate learners' (successful) uptake of formulaic forms and compared the findings with the variables that predict learners' (successful) uptake of non-formulaic forms. The investigation of variables predicting learners' successful uptake of FSs is important given that learners' production of successful uptake has been associated with their noticing target forms and L2 learning (Egi, 2010; Loewen, 2004). Therefore, this study examined incidental FonF characteristics that mediate learners' provision of (successful) uptake following formulaic vs. non-formulaic forms through the following research questions:

1. How often does (successful) uptake occur in incidental focus-on-form episodes (FFEs) with formulaic vs. non-formulaic foci in three advanced English as foreign language (EFL) classes?

2. What characteristics of incidental FonF best predict learners' production of (successful) uptake in FFEs with formulaic vs. non-formulaic foci?

3. Method

3.1. Study context

This study was conducted in an intensive adult EFL program in a language school in Urmia, Iran. The participants included a total of six teachers (M = 31.5, SD = 3.2) teaching six intact classes, and 68 learners (M = 27.3, SD = 3.7). The learners were of Farsi (n = 19), Azeri (n = 36), and Kurdish (n = 13) language backgrounds. The language school offers general English classes at all proficiency levels. The learners had three to nine years of the language learning experience. Data were collected from six advanced level classes from different sections of the same course with the same

textbook and syllabus. The learners were placed in the advanced level classes based on their scores on an IELTS test created in house and administered by the school officials. The learners' mean scores on sections of the IELTS placement test (with 0-9 band score range) were 8.6 in listening, 8.2 in reading, 8 in writing, and 7.6 in speaking. Based on the researcher's evaluation of the classroom discourse using the American Council on the Teaching of Foreign Languages proficiency guidelines (2012) and learners' test scores on the placement IELTS test, the learners were assessed to be approximately at Advanced-low to Advanced-mid proficiency levels.

Six classes were taught by six EFL teachers for whom English is an L2. Their teaching experience ranged from five to 17 years. The teachers held bachelor's, master's, and doctorate degrees in TEFL. TEFL degree programs in Iran are university-level programs offering general English courses and disciplinary courses on applied linguistics, second language acquisition, language assessment, and research methods in these domains. With no experience living or teaching in an English-speaking country, the teachers taught EFL in public and private schools at different proficiency levels. Unlike the public schooling system in Iran that aims to prepare students for the discrete-point university entrance exam, private language schools are expected to implement communicative language teaching. The teachers integrated the language skills of listening, speaking, reading, and writing employed communicative tasks including role-plays, information-gap tasks, opinion-gap tasks, etc. Teachers engaged learners in pair and group work, created opportunities for discussion of various topics, used prompt-based writing and speaking activities and game-based activities. Both teachers and learners were asked to complete consent forms.

3.2. Procedure

The data included 36 hours of audio recordings from six intact advanced adult EFL classes (six hours per class). The audio recordings were captured with a digital wireless voice recorder in each classroom. The teachers were asked to wear a wireless voice recorder with a clip-on microphone. The corpus used in this study was comprised of verbal teacher-learner interactions in one-on-one, small group, and whole-class. All six teachers used the same textbook (*Speakout*, Clare & Wilson, 2016), syllabus, and supplementary materials. The teachers were asked to teach their normal classes and were not informed of the study's focus. Five hours of communicative-oriented interactions from each class were analyzed with a total of 30 hours of data after excluding the time allotted for roll call, greeting, and teaching isolated linguistic structures.

3.2.1. Coding FonF episodes

Following other studies (Ellis et al., 2001), the unit of analysis was a focus-on-form episode (FFE) in this study. An FFE is defined as "the discourse from the point where the attention to linguistic form starts to the point where it ends, due to a change in topic back to message or sometimes another focus on form" (Ellis et al., 2001: 294). The beginning of an FFE is marked by a learner's non-target-like use of a form in reactive FonF, a learner's query about a form in student-initiated FonF, and the teacher's query or comment about a form in teacher-initiated FonF (Ellis et al., 2001). The end of an FFE is signaled by learner uptake, topic continuation by the teacher or learner, or another FFE (Ellis et al., 2001).

The researcher (the first coder) and a research assistant (the second coder) were involved in all coding procedures. The inter-coder reliabilities were achieved using Cohen's (1960) Kappa measure of agreement. Kappa values are reported and marked as κ in the respective tables and appendices. Kappa values between 0.81–1.00 are considered as high reliability (Cohen, 1960). The two coders independently listened to the audio recordings from one session of the class and identified FFEs. The inter-coder reliability was found to be high, $\kappa = .92$. The coders established full reliability in all coding procedures by resolving any discrepancies in coding. The first coder listened to all audio recordings, identified all instances of FFEs, and transcribed them. Table 1 shows the linguistic foci of FFEs. Table 2 presents FFEs with different foci and characteristics.

Table 1. Linguistic categories

Linguistic focus		Description Reliability κ = .87
A. Formulaic forms	1. Collocation	Linguistic features with formulaic nature. A lexical collocation consists of two content words (i.e., adjective, adverb, noun, or verb) that both contribute almost equally to its entire meaning. A grammatical collocation consists of a dominant content word (i.e., a noun, a verb, or an adjective) and a subordinate grammatical structure (i.e., a preposition, an infinitive, or a clause) (Benson, Benson, & Ilson, 2010).
	2. Lexical bundle	Lexical bundles are recurrent expressions that "commonly go together in natural discourse" (Biber, Johansson, Leech, Conrad, & Finegan, 1999: 990).
	3. Idiom	Idioms are "opaque invariant word combinations" (Warren, 2005: 35).
	4. Compound	Compounding is the creation of a word with a specific meaning by blending two existing words (Wood, 2020).
B. Non-formulaic forms	1. Grammar	Linguistic features with non-formulaic nature. Subject-verb agreement, tense, plurals, word order, question formation, negation, determiners, pronouns, prepositions, plural, verb morphology, sentence construction, etc
	2. Pronunciation	Segmental and supra-segmental aspects of the phonological system that are not related to bound grammatical morphemes. Pronunciation of words.
	3. Vocabulary	Meaning of single-word items.
	4. Spelling	The orthographic form of words (Ellis et al., 2001).

Table 2. Instances of FFEs with different characteristics

Example 1: Episode with formulaic focus	**Characteristics**	**Category**
1 S Rush hour means?	Type	Student-initiated
2 T The busy hour (..) I mean (.) for example, from 5:30 to 6 in the evening (.)	Linguistic focus Source	Lexical collocation Message
3 S Yes!	Complexity	Complex
4 T In Ramadan (.) it is the rush hour (.) people are rushing home to break their fast as quickly as possible. ha!	Directness Emphasis Timing	Direct Heavy Immediate
5 S Yeah (.) I drive home in the rush hour.	Response Uptake	Provide Uptake, successful
Example 2: Episode with formulaic focus	**Characteristics**	**Category**
1 S On other words (.)	Type	Reactive
2 T **IN other words** (.)	Linguistic focus	Lexical bundle
3 S Ahh (.) sorry (.) in other words (.) some people don't care about the dangers of smoking	Source Complexity Directness Emphasis Timing Response Uptake	Code Simple Direct Light Immediate Provide Uptake, successful
Example 3: Episode with formulaic focus	**Characteristics**	**Category**
1 T What does **tie the knot** mean? initiated any ideas?	Type Linguistic focus	Teacher-initiated Idiom
2 S گره زدن (*the equivalent of to tie in Persian*) /'gereh zædæn/	Source Complexity	Message Complex
3 Ss @	Directness	Direct
4 T NO (.) It's an expression.	Emphasis	Heavy
5 T For example (.) my fiancé and I are going to tie the knot (..) means? We'll get married. To tie the knot (.) means to get married. OK, back to the reading …	Timing Response Uptake	Immediate Provide No opportunity

Example 4: Episode with formulaic focus	**Characteristics**	**Category**
1 S Teenagers today (.) don't know how to deal for their problems.	Type Linguistic focus	Reactive Grammatical collocation
2 T deal for or **deal WITH?**	Source	Code
3 S Deal with.	Complexity	Simple
4 T Right!	Directness	Indirect
5 S Thanks (.) dealing with the …	Emphasis Timing Response Uptake	Light Immediate Elicit Uptake, successful
Example 5: Episode with formulaic focus	**Characteristics**	**Category**
1 S I want to travel to other countries (..) What we say? What hiking?	Type Linguistic focus	Student-initiated Compound
2 T **Hitchhiking** (.) you mean?	Source	Message
3 Like you don't have a car, but you want to travel to another city, and you thumb *(teacher demonstrating the thumb signal for hitchhiking).* Well (.) class is over.	Complexity Directness Emphasis Timing Response Uptake	Complex Direct Heavy Immediate Provide No opportunity
Example 6: Episode with non-formulaic focus	**Characteristics**	**Category**
1 S She influence my entire life	Type	Reactive
2 T She **influence** my entire life INFLUENCE?	Linguistic focus Source	Grammar Code
3 S Yeah.	Complexity Directness Emphasis Timing Response Uptake	Simple Direct Light Delayed Elicit Uptake, successful

Example 7: Episode with non-formulaic focus	**Characteristics**	**Category**
1 S How do you pronounce this?	Type	Student-initiated
2 T /nuːˈmoʊniə/	Linguistic focus	Pronunciation
	Source	Code
	Complexity	Simple
	Directness	Direct
	Emphasis	Light
	Timing	Immediate
	Response	Elicit
	Uptake	No uptake
Example 8: Episode with non-formulaic focus	**Characteristics**	**Category**
1 T Did you get the meaning of **sabotage**?	Type	Teacher-initiated
2 Ss (....)	Linguistic focus	Vocabulary
3 T To destroy. To damage.	Source	Message
	Complexity	Simple
	Directness	Direct
	Emphasis	Light
	Timing	Immediate
	Response	Provide
	Uptake	No uptake

3.2.2. Coding foci of FFEs

Chomsky (1965) posited that depending on the context, the word sequence "*decide on a boat*" in the sense of "*decide while on a boat*" is a loose non-formulaic construction. In contrast, the same word sequence in the sense of "*decide what boat to choose*" is a close formulaic construction. The dual-nature view of language is substantiated by Sinclair's (1991) open-choice principle (non-formulaic forms) and the idiom principle (formulaic forms).

Formulaic constructions (e.g., *on the other hand*) are distinct from non-formulaic constructions (e.g., *on the other foot*) in that in the former, at least one constituent cannot be replaced by a synonymous word or phrase without changing function, meaning, or idiomaticity, which is known as restricted exchangeability (Erman & Warren, 2000). Formulaic forms are characterized as "any sequence of two or more words that are perceived to be more constrained than usual in their co-occurrence" (Hudson & Wiktorsson, 2009: 81). Non-formulaic forms refer to "elements used in

their literal senses and freely substitutable" (Howarth, 1998: 28). Non-formulaic forms are generated by syntactic analysis in which "the only restraint is grammaticalness" (Sinclair, 1991: 109).

Previous studies made a distinction between non-formulaic (*drop the beans*) and formulaic (*spill the beans*) language, albeit for different research purposes (e.g., Gholami, 2021a, 2021b, 2021c; Gholami, Karimi, & Atai, 2017). For the purpose of this study, non-formulaic forms were differentiated from formulaic forms. Formulaic forms were analyzed as a separate category in this study as FSs "are nevertheless significant enough to be the focus of research, and a theoretical category meriting particular attention" (Buerki, 2016: 15). FSs were not examined in terms of their individual components as they "constitute single choices, even though they might appear to be analyzable into segments" (Sinclair, 1991: 110). Similarly, FSs were not subsumed under the category of vocabulary because they "are not the products of general rules applying to words, and nor do they, in general, behave like single words" (Buerki, 2016: 16).

The two coders marked FFEs for formulaic and non-formulaic foci. Following Wood's (2020) suggestion to rely on native-speaker intuition, the second coder, a native speaker of American English (a Ph.D. in applied linguistics), was involved in identifying FFEs with formulaic foci. The coders read Wood's (2020) summary of checklists employed for the identification of FSs. Based on the speaker-external view of formulaicity, FSs are demarcated from non-formulaic forms in terms of their formal properties, that is, semantic irregularity (*kick the bucket*), syntactic idiosyncrasy (*by and large*), pragmatic functions (*what's up?*), or frequency of co-occurrence (*rock and roll*) (Myles & Cordier, 2017). The coders used Wray and Namba's (2003) checklist (Appendix A) as the main checklist for judging formulaic vs. non-formulaic foci of FFEs. Their checklist is comprehensive with 11 criteria and guidelines on how to employ different criteria for various datasets involving error-free, error in the form, and error in usage (See Wray & Namba, 2003, for detailed guidelines).

After dividing FFEs into episodes with formulaic and non-formulaic foci, they were coded for categories of the (non)formulaic forms presented in Table 1. Zhao & Bitchener (2007: 438) provide sub-categories of non-formulaic forms. In FFEs with formulaic foci, FSs are focused on to address: (a) form, meaning, or usage; (b) lexical selection; and/or (c) lexical formation (Millar, 2011) (Appendix B). In the FS "*a piece of cake*" (easy to do), the figurative sense is lost by lexical misselection "*a piece of pancake*" or lexical misformation "*the piece of cake*" (Xu, 2015). Nesselhauf (2005) delimited formulaic errors (i.e., those stemming directly from the phraseological status of FSs) from non-formulaic errors (i.e., those about the syntactic rules governing non-formulaic constructions). For instance, in the word sequence "*she make judgments,*" if the subject-verb agreement was targeted, the FFE was marked as non-formulaic. Nevertheless,

if lexical selection (*pass judgment)* or lexical formation (*judgment* is never pluralized because of the formulaic status) were targeted, the FFE was marked as formulaic.

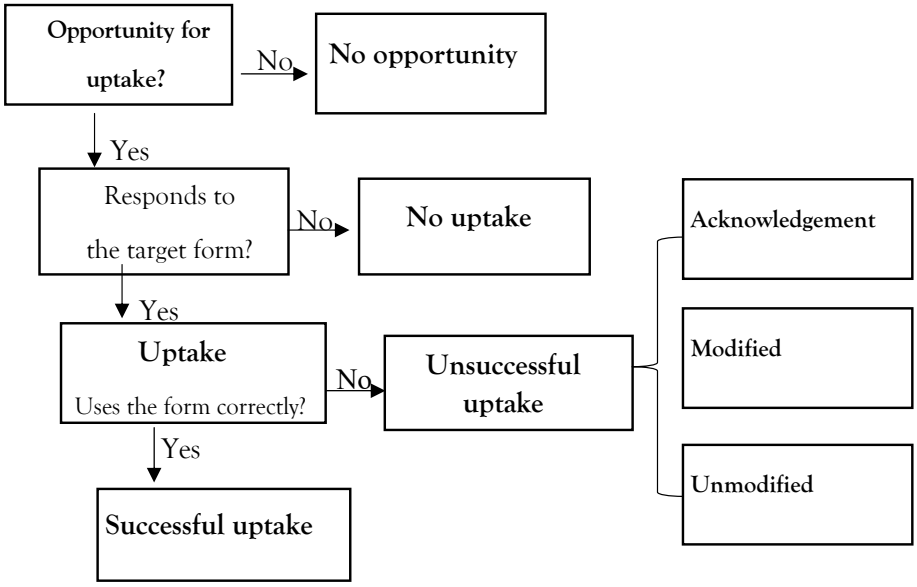
3.2.3. Coding categories of FSs

Collocations are distinct from non-formulaic combinations due to *the arbitrary restriction on combinability* (Nesselhauf, 2005). For instance, the word string "*kick the stone*" is a free combination as all its components could be replaced by other words (*kick the ball*). However, in the word string "*kick a habit*," the base (*habit*) in the sense used in the collocational unit could be combined with other linguistic items, but the collocator (*kick*) in the sense used in the collocational unit is arbitrarily restricted to certain elements (*break a habit*) (Nesselhauf, 2005). Two major sub-categories of collocations include lexical collocations (*brain drain*) and grammatical collocations (*on purpose*) (Xu, 2015) that were coded using the schemes presented in Appendices C and D, respectively. Following Nesselhauf (2005), the identified collocations were crosschecked with *The BBI Combinatory Dictionary of English* (2010). Lexical bundles (*the extent to which*) were coded using Simpson-Vlach & Ellis's (2010) *Academic Formulas List* (AFL), which provides lists of lexical bundles frequently found in English written and spoken discourses (See Appendix E). Idioms were identified using the *Oxford Dictionary of English Idioms* (2010) and *The Oxford Dictionary of Idioms* (2005). Compounds were coded using the coding scheme presented in Appendix F.

3.2.4. Coding uptake

As Figure 1 illustrates, FFEs were coded for: (a) *no opportunity for uptake* when a learner has no chance to react to target form because the teacher or another learner immediately continues a topic; (b) *no uptake* when the learner does not react to target form despite having a chance to react; (c) *uptake* when the learner responds to target form triggering the FFE. The inter-coder reliability for uptake was found to be high, $\kappa = 0.85$. FFEs with uptake were further coded for (d) *successful uptake* when the learner successfully incorporates target form into production by rephrasing, correcting the error, or using the target form correctly in an example; (e) *unsuccessful uptake* when the learner's response to the target form requires further correction. Unsuccessful uptake involves *acknowledgment* when learner acknowledges the recognition of target form by uttering *thanks*, *yeah,* etc.; *modified* when learner modifies the error incorrectly or partially correctly; and *unmodified* when the learner does not modify the error, expresses difficulty understanding the target form, or circumvents the use of target form. The inter-coder reliability for successful uptake was found to be high, $\kappa = 0.82$. Learners' errors with forms that were not targeted in an FFE were disregarded.

Figure 1. Coding scheme of uptake (adopted from Egi, 2010).



3.2.5. Coding characteristics of FFEs

All FFEs were further coded for the characteristics presented in Table 3. Table 2 presents instances of FFEs with different characteristics.

Table 3. Characteristics of FFEs (adopted from Loewen, 2005: 376)

Characteristic	Definition	Categories	k = Reliability
Type	Instigation	*Reactive FFE:* It "arises when learners produce an utterance containing an actual or perceived error, which is then addressed usually by the teacher but sometimes by another learner. Thus, it supplies learners with negative evidence" (Ellis et al., 2001: 413). *Preemptive FFE:* It involves a learner or the teacher "initiating attention to form even though no actual problem in production has arisen" by raising a query (Ellis et al., 2001: 414).	*k* = 0.855
Source	Reason for instigation	*Code:* An FFE involves the inaccurate use of a form with no apparent miscommunication (i.e., negotiation of the form). *Message:* An FFE involves problems with understanding meaning (i.e., negotiation of the meaning).	*k* = 0.857
Complexity	Length	*Simple:* An FFE involves a single exchange to resolve the linguistic problem. *Complex*: An FFE involves two or more exchanges to resolve the linguistic problem	*k* = 0.789
Directness	Explicitness	*Indirect:* Attention to form is implicit. *Direct:* Attention to form is explicit.	*k* = 0.832
Emphasis	Complexity + directness	Light: An FFE is indirect and simple *Heavy*: An FFE is direct and/or complex	*k* = 0.790
Timing	Response timing	Immediate: An FFE is immediately provided Delayed: An FFEs is provided with some delay.	k = 0.986
Response	Type of feedback	*Provide:* The teacher provides learners with target form through recasts or explicit corrections. *Elicit*: The teacher elicits target form from learners through clarification requests, repetitions, metalinguistic clues and elicitations	k = 0.809
Uptake	Student response to feedback	*Uptake:* The learner responds to the target form *No uptake:* The learner does not respond to target form.	k = 0.991
Successful uptake	Quality of student response	*Successful uptake:* The learner incorporates the target form into production. *Unsuccessful uptake:* The learner does not successfully incorporate target form into production.	k = 0.889

Note. *k* refers to the inter-coder reliability.

4. Results

A total of 1,425 FFEs were identified in 30 hours of teacher-learner interactions. The focus of 551 episodes was formulaic, and 874 FFEs focused on non-formulaic forms. Therefore, learners' attention was shifted to non-formulaic forms (61%) more often than FSs (39%). Table 4 shows (successful) uptake rate in FFEs with (non)formulaic foci. Instances of FFEs where there was no opportunity for learners to produce uptake were excluded (n = 63). Uptake occurred in a total of 746 FFEs comprising more than half of the total episodes (55%). Chi-square tests were conducted to investigate the statistical associations between (non)formulaic foci of FFEs and (successful) uptake rate. The inferential statistics were calculated using Statistical Package for the Social Sciences (SPSS) 25.0. An alpha level of $p < 0.05$ was set for all chi-square tests. The association between the (non)formulaic foci of FFEs and uptake rate was found to be significant, χ^2 (1, 1362) = 131.746, p = 0.000. The effect size was medium, w = 0.3. Therefore, learners produced uptake more often in FFEs with formulaic foci (52%) than non-formulaic foci (48%). The association between (non)formulaic foci of FFEs and successful uptake rate was found to be significant, χ^2 (1,746) = 100.684, p = 0.000. The effect size was medium, w = 0.4. Hence, learners produced successful uptake more often in FFEs with formulaic foci (66%) than non-formulaic foci (34%).

Table 4. (Successful) uptake rate according to (non) formulaic foci

	No uptake		Uptake		Successful uptake		Unsuccessful uptake	
	No.	%	No.	%	No.	%	No.	%
Formulaic	134	22	389	52	314	66	75	28
Non-formulaic	482	78	357	48	162	34	195	72
Total	616	45	746	55	476	64	270	36

Note. The percentages are rounded off.

Binary logistic regression analyses were performed to investigate the characteristics that mediate the occurrence of (successful) uptake in FFEs with formulaic vs. non-formulaic foci. The logistic regression analysis results are reported in terms of odds ratio and 95% confidence interval for each independent variable (Hosmer & Lemeshow, 2000). The logistic regression analysis for uptake in FFEs with formulaic foci revealed an overall percentage accuracy of 68.6%. The variables of type, source, complexity, emphasis, timing, and response were found to be significant predictors of uptake in FFEs with formulaic foci. As Table 5 shows, the odds ratio for the type of incidental FonF was found to be 3.841, meaning that preemptive FFEs targeting FSs were roughly

four times more likely to lead to uptake than reactive episodes. The odds ratio for the source of FonF was found to be 14.889, indicating that message-related FFEs with formulaic foci were approximately fifteen times more likely to contain uptake than code-related episodes. Complexity has an odds ratio of 6.099, indicating that FFEs with formulaic foci involving many uptake moves were six times more likely to result in uptake than simple episodes. Emphasis has an odds ratio of 4.943, meaning that FFEs with formulaic foci involving several direct uptake moves stand roughly five times more chance of culminating in uptake than simple indirect episodes. The odds ratio for timing was 0.206, meaning that delayed FFEs with formulaic foci were one-fifth as likely to contain uptake as immediate episodes. The odds ratio for the variable of response was found to be 4.948, meaning that eliciting target forms from learners was roughly five times more likely to lead to an uptake in FFEs with formulaic foci than providing learners with correct forms.

Table 5. Logistic regression for uptake in formulaic FFEs

			95% confidence intervals		
Variable	S.E.	Odds ratio	Lower	Upper	Significance
Type	.235	3.841	2.422	6.09	.000
Source	.286	14.889	8.507	26.058	.000
Complexity	.247	6.099	3.757	9.900	.000
Emphasis	.239	4.943	3.097	7.892	.000
Timing	.238	206	129	328	.000
Response	.236	4.948	3.117	7.854	.000

The second logistic regression analysis conducted for successful uptake in FFEs with formulaic foci revealed an overall percentage accuracy of 74.1%. The variables of type, source, complexity, emphasis, timing, and response were found to be significant predictors of successful uptake in FFEs with formulaic foci. As Table 6 shows, preemptive FFEs with formulaic focus were two and a half times (odds ratio = 2.547) more likely to lead to successful uptake than reactive episodes. Message-related FFEs with formulaic foci were seven times (odds ratio = 7.386) more likely to culminate in successful uptake than code-related ones. The odds ratio for complexity was found to be 2.124, meaning that complex FFEs with formulaic foci were twice more likely to result in successful uptake than simple ones. Heavy emphasis (odds ratio = 4.545) on FSs in FFEs with formulaic foci involving several direct uptake moves was four and a half times more likely to contain successful uptake than light emphasis involving simple indirect uptake moves. Delayed FFEs with formulaic foci (odds ratio = 0.196) were

roughly one-fifth as likely to culminate in successful uptake as immediate episodes. Finally, eliciting target forms from learners was almost five times (odds ratio = 4.946) more likely to lead to successful uptake in FFEs with formulaic foci than providing learners with correct forms.

Table 6. Logistic regression for successful uptake in formulaic FFEs

			95% confidence intervals		
Variable	S.E.	Odds ratio	Lower	Upper	Significance
Type	.286	2.547	1.455	4.461	.001
Source	.313	7.386	4.002	13.632	.000
Complexity	.288	2.124	1.208	3.736	009
Emphasis	.296	4.545	2.544	8.120	.000
Timing	.299	196	109	353	.000
Response	.300	4.946	2.749	8.899	.000

The logistic regression analysis for uptake in FFEs with non-formulaic foci showed an overall percentage accuracy of 61.5%. The variables of type, source, complexity, directness, emphasis, timing, and response were found to be significant predictors of uptake in FFEs with non-formulaic foci. As Table 7 presents, the odds ratio of type was 0.259, meaning that preemptive FFEs with non-formulaic foci were one-third as likely to contain uptake as reactive episodes. The source's odds ratio was 0.277, meaning that message-related FFEs with non-formulaic foci were one-third as likely to contain uptake as code-related episodes. The odds ratio of complexity was 2.049, indicating that complex FFEs with non-formulaic foci were twice more likely to end with uptake than simple episodes. Directness has an odds ratio of 0.304, meaning that indirect episodes with non-formulaic foci were one-third as likely to culminate in successful uptake as direct episodes. The odds ratio of emphasis was 2.049, indicating that FFEs with non-formulaic foci involving heavy emphasis were twice more likely to result in uptake than episodes with light emphasis. The odds ratio for timing was found to be 4.628, indicating that delayed FFEs with non-formulaic foci were roughly five times more likely to end with uptake than immediate episodes. The response has an odds ratio of 0.476, indicating that eliciting target forms from learners in FFEs with non-formulaic foci was roughly one-fifth as likely to contain uptake as providing learners with target form.

Table 7. Logistic regression for uptake in non-formulaic FFEs

			95% confidence intervals		
Variable	S.E.	Odds ratio	Lower	Upper	Significance
Type	.150	.259	.199	.358	.000
Source	.153	.277	.205	.374	.000
Complexity	.155	2.049	1.513	2.774	.000
Directness	.152	.304	.226	.409	.000
Emphasis	.155	2.049	1.513	2.774	.000
Timing	.158	4.628	3.396	6.306	.000
Response	.148	.476	.357	.637	.000

The logistic regression analysis for successful uptake in FFEs with non-formulaic foci showed an overall percentage accuracy of 54%. The variables of type, source, complexity, directness, emphasis, timing, and response were found to be significant predictors of successful uptake in FFEs with non-formulaic focus. As presented in Table 8, preemptive FFEs with non-formulaic foci were one-third (odds ratio = 0.265) as likely to contain successful uptake as reactive ones. Message-related FFEs with non-formulaic foci were roughly one-fourth (odds ratio = 0.396) as likely to result in successful uptake as code-related episodes. Complex FFEs with non-formulaic foci were three times (odds ratio = 3.236) more likely to end with successful uptake than simple episodes. Indirect FFEs with non-formulaic foci were one-third (odds ratio = 0.347) as likely to contain successful uptake as direct episodes. FFEs with non-formulaic foci involving several direct moves were three times (odds ratio = 3.152) more likely to lead to successful uptake than those with light focus. Delayed FFEs with non-formulaic foci were five times (odds ratio = 5.188) more likely to result in successful uptake than immediate ones. Finally, eliciting target forms from learners in FFEs with non-formulaic foci was one-third (odds ratio = 0.297) as likely to culminate in successful uptake than providing learners with target forms. Therefore, the findings of logistic regression analyses showed that FonF characteristics predicting learners' production of (successful) uptake were different for formulaic vs. non-formulaic forms targeted in FFEs.

Table 8. Logistic regression for successful uptake in non-formulaic FFEs

			95% confidence intervals		
Variable	S.E.	Odds ratio	Lower	Upper	Significance
Type	.253	.265	.195	.354	.000
Source	.236	.396	.250	.629	.000
Complexity	.239	3.236	2.028	5.165	.000
Directness	.237	.347	.218	.552	.000
Emphasis	.238	3.152	1.977	5.927	.000
Timing	.250	5.188	3.181	8.461	.000
Response	.240	.297	.185	.474	.000

5. Discussion

5.1. Uptake of (non)formulaic forms

This study examined the extent to which learners produced (successful) uptake following FSs targeted in FEEs compared to non-formulaic forms targeted in FFEs. Also, incidental FonF characteristics that mediated learners' production of (successful) uptake in FFEs following target FSs were compared with target non-formulaic forms. The findings indicated that non-formulaic forms (61%) were more frequently focused on than formulaic forms (39%) in the observed classrooms. However, learners produced uptake (52%) and successful uptake (66%) more often when FSs were targeted in FFEs than non-formulaic forms (uptake = 48%, successful uptake = 34%). In line with this study's findings, Lyster (2001) found that learners generated (successful) uptake more often following lexical items than grammar and pronunciation. Learners' production of successful uptake following FonF has been associated with their subsequent language learning (e.g., Egi, 2010; Ellis et al., 2001; Loewen, 2004; Lyster, 2001). Lyster (2001) concluded that FonF directed at lexically-oriented items results in a higher successful uptake rate and subsequently effective learning than FonF targeting morphosyntax.

Learners' higher production of (successful) uptake following FSs in FFEs than non-formulaic forms as found in this study is attributed to the greater saliency, communicative value, and noticeability of FSs. Wulff (2019) posits that vocabulary items are more salient than morphosyntax, and FSs are even more salient than single-word vocabulary items. In addition, FSs perform different discourse functions and play a key role in communicating meaning (Wullf, 2019). In this regard, Kecskes (2016) notes that FSs are "not only salient but also functionally significant" (p. 10). Learners' production of successful uptake is deemed an indicator of noticing target forms (Egi, 2010). Therefore, the higher proportion of successful uptake following target FSs as found in this study suggests the greater noticeability of FSs than non-formulaic forms. Li & Vuono (2019) maintain that the nature of linguistic forms influences the degree of noticeability of forms by learners and their subsequent production of successful uptake. This study's findings support the conviction that forms with formulaic nature are more salient, noticeable, and communicatively more important than forms involving grammar, pronunciation, and single-word vocabulary as indicated by markedly higher (successful) uptake rate following FSs.

In this study, learners' attention was raised to non-formulaic forms (61%) more often than FSs (39%) in FFEs. Nonetheless, learners tended to produce more successful uptake following target FSs (66%) than no-formulaic forms (34%). While not questioning targeting non-formulaic forms, this study suggests that learners' attention be drawn to formulaic aspects of language more frequently through FFEs to achieve a balanced representation of formulaic and non-formulaic aspects of language in incidental FonF. Learners' production of successful uptake in incidental FonF provides them with the opportunity "to notice linguistic items and structures at the very moment they are most needed for communication" (Loewen & Sato, 2019: 10), leading to restructuring interlanguage toward target models (Panova & Lyster, 2002). Therefore, another line of this study examined incidental FonF characteristics that mediate learners' production of (successful) uptake following target FSs compared to non-formulaic forms in FFEs.

5.2. FonF characteristics predicting learner uptake

This study's findings showed that FonF characteristics that predicted learners' production of (successful) uptake were different for formulaic forms vs. non-formulaic forms targeted in FFEs. To the best of the researcher's knowledge, to date, no studies have examined incidental FonF characteristics that influence learners' production of (successful) uptake following FSs to make comparisons with the findings of this study. In this study, preemptive FonF was a strong predictor of learners' production of (successful) uptake when FSs were targeted in FFEs. In contrast, reactive FonF was

a strong predictor of learners' production of (successful) uptake when non-formulaic forms were targeted in FFEs.

A partial explanation for differences in these findings might lie in the formulaic vs. non-formulaic nature of target forms. Ellis et al. (2001) differentiate between actual and perceived linguistic gaps in learners' knowledge. Student-initiated preemptive FonF, where learners take responsibility for their own learning by raising queries, has the potential to address learners' actual linguistic gaps (Ellis et al., 2001). However, teacher-initiated reactive FonF, where teachers provide corrective feedback to learners' errors, may address their perceived errors (Ellis et al., 2001). The learners in this study were advanced adult learners taking tuition-based classes, so they might have taken responsibility for addressing their problems with FSs by using student-initiated preemptive FonF and subsequently producing (successful) uptake to communicate their meaning. In this regard, Loewen (2004) notes that instructional context and learners' proficiency level, among other factors, influence learners' use of student-initiated FonF and the production of successful uptake.

On the other hand, learners' errors with non-formulaic forms involving grammar at advanced level classes might have arisen due to the spontaneous language use as advanced learners are expected to have already developed substantial knowledge of grammar and vocabulary (Meunier, 2012), which could result in teachers' using reactive FonF to correct learners' errors, leading to learners' higher successful uptake of non-formulaic forms following reactive FonF. Therefore, this study's findings suggest that to increase learners' production of successful uptake in FFEs, teachers should use reactive FonF to address non-formulaic forms and provide opportunities for learners to take the initiative and preemptively address FSs through preemptive FonF.

Another variable predicting learners' successful uptake in FFEs with formulaic vs. non-formulaic foci was the type of corrective feedback in FonF. The findings showed that eliciting target FSs from learners through output-prompting feedback types strongly predicted learners' production of (successful) uptake in FFEs. However, providing learners with target non-formulaic forms through input-providing feedback types strongly predicted learners' production of (successful) uptake in FFEs. Feedback strategies are categorized into input-providing (i.e., using recasts and explicit corrections) and output-prompting types (i.e., eliciting target forms, making clarification requests, repeating errors, and providing metalinguistic clues) (Li & Vuono, 2019).

The findings of this study suggest that the nature of forms in terms of formulaic vs. non-formulaic mediate the effectiveness of different corrective feedback types as measured by successful uptake rate. Likewise, the findings of Brown's (2016) meta-analysis demonstrated that the nature of linguistic target moderates the efficacy of

corrective feedback. Similarly, Lyster & Mori's (2006) counterbalance hypothesis indicates that the provision of corrective feedback should vary depending on instruction's focus on morphosyntax or lexical items. Additionally, the nature of linguistic target in terms of whether it entails item learning (FSs) or system learning (grammar) is a factor that "can determine the relative effect of different strategies on noticing, uptake with repair, and acquisition" (Ellis, 2017: 11). Ellis (2017) argues that teachers should not select corrective feedback types randomly "but should apply them systematically" (p. 12). Hence, to promote learners' production of successful uptake following FSs, this study's findings suggest using output-prompting corrective feedback types. However, to promote learners' production of successful uptake following non-formulaic forms, this study's findings suggest using input-providing corrective feedback types.

Source of FonF was another variable mediating learners' provision of (successful) uptake following FSs vs. non-formulaic forms in FFEs. Learners produced more (successful) uptake following FSs when FFEs involved message conveyance. However, learners produced more (successful) uptake following non-formulaic forms when FFEs involved the accuracy of form. These differences in findings could be explained by the distinction between the negotiation of form versus negotiation of meaning, as Ellis et al. (2001) suggested. Ellis et al. (2001) argue that forms pertaining to phonology and morphosyntax mainly require negotiation of form to enhance linguistic accuracy despite the lack of any communication breakdown. On the other hand, lexically-oriented forms, including FSs require negotiation of meaning to resolve meaning and communication problems (Ellis et al., 2001). Hence, FSs lend themselves more to the negotiation of meaning, resulting in a higher (successful) uptake rate. In contrast, non-formulaic forms lend themselves more to the negotiation of form, resulting in a higher (successful) uptake rate.

Timing of FonF was another variable mediating learners' production of (successful) uptake following target FSs vs. non-formulaic forms in FFEs. The immediate focus on FSs was a strong predictor of (successful) uptake when FSs where targeted in FFEs. On the contrary, the delayed focus on non-formulaic forms was a strong predictor of (successful) uptake when non-formulaic forms were targeted in FFEs. There are inconsistent findings on the efficacy of immediate and delayed FonF (e.g., Li, Zhu, & Ellis, 2016; Quinn & Nakata, 2017). Quinn & Nakata (2017) argue that delayed FonF is effective in L2 development according to the distributed practice effect, indicating that longer intervals between attention to a form could lead to more effective long-term retention than shorter or no intervals. In contrast, Doughty (2001) posits that FonF is optimal when it occurs within 60 seconds of the trigger when humans can maintain active mental representations in the working memory. This study's findings suggest

that the nature of forms in terms of formulaic vs. non-formulaic nature mediates the effectiveness of delayed vs. immediate FonF as measured by the successful uptake rate. Therefore, to promote learners' production of (successful) uptake, it is suggested that teachers use immediate and delayed FonF to address FSs and non-formulaic forms in incidental FonF, respectively.

Finally, the variables of complexity, directness, and emphasis were found to predict learners' production of (successful uptake) following FSs and non-formulaic forms in the same manner. Therefore, the findings of this study showed that some of FonF characteristics that predicted learners' production of (successful) uptake were different for formulaic vs. non-formulaic forms targeted in FFEs.

6. Conclusion

The present study investigated learners' production of (successful) uptake following FSs vs. non-formulaic forms in incidental FonF. Moreover, incidental FonF characteristics predicting learners' (successful) uptake of FSs vs. non-formulaic forms targeted in FFEs were examined. The findings showed that learners' attention was shifted to non-formulaic forms (61%) more often than FSs (39%) in FFEs. Nevertheless, learners produced successful uptake more often when FSs (66%) were targeted in FFEs than non-formulaic forms (34%). The higher rate of successful uptake following FSs could be ascribed to the greater saliency, noticeability, and communicative value of FSs (Wulff, 2019). In this study, teachers raised learners' attention to non-formulaic forms strikingly more often than FSs, while learners produced successful uptake markedly more often following FSs owing to the importance of FSs for communication. Therefore, in the light of these findings, it is suggested that teachers consider FSs as legitimate and essential targets among other linguistic forms and use incidental FonF opportunities to draw learners' attention to FSs as frequently as non-formulaic forms. The findings of this study suggest that teacher professional development programs raise pre-service and practicing teachers' awareness of the significance of FSs in communicative language use and the importance of incorporating FSs into incidental FonF practices.

Furthermore, the literature on formulaic language indicates that even advanced L2 learners are at a disadvantage to learn FSs due to the traditional emphasis on grammar and vocabulary in curriculum designs, materials development, and L2 pedagogy and assessment (Meunier, 2012; Wray, 2019). Given the significance of FSs for learners as indicated by their markedly higher successful uptake of FSs found in this study, it is suggested that FSs gain prominence in curriculum designs for L2 pedagogy as

advocated by the proponents of chunk-oriented pedagogy (Meunier, 2012). Meunier (2012) attributed learners' inadequate knowledge of formulaic language compared to their knowledge of grammar and vocabulary partly to the paucity of corpus-informed materials. Corpus findings indicate that FSs are ubiquitous in language use (Erman & Warren, 2000). Therefore, given the importance of FSs for learners in communication, as suggested by the higher successful uptake of FSs, it is recommended that a rich repertoire of FSs should be incorporated into materials developed for L2 pedagogy in line with corpus findings.

Another line of this study examined incidental FonF characteristics predicting learners' (successful) uptake of FSs vs. non-formulaic forms targeted in FFEs. The findings of logistic regression analyses revealed that incidental FonF characteristics in terms of the type of incidental FonF (reactive vs. preemptive FonF), timing of FonF (immediate vs. delayed), the source of FonF (problem with communication vs. problem with the accuracy of the form), and type of corrective feedback (output-prompting vs. input-providing) varied in predicting learners' production of (successful) uptake following FSs vs. non-formulaic forms. Immediate preemptive focus on FSs to address learners' problem with meaning rather than form and eliciting target FSs from learners through output-prompting feedback types were strong predictors of learners' uptake of FSs and successfulness of it. In contrast, delayed reactive focus on non-formulaic forms to address learners' problem with form rather than meaning and providing target non-formulaic forms for learners through input-providing feedback types were strong predictors of learners' uptake of non-formulaic forms and successfulness of it.

Hence, the findings of this study lend support to Loewen's (2005) conviction that "different kinds of focus on form might be needed for different aspects of language" (p. 382). Likewise, DeKeyser (2012) notes that different linguistic forms call for different needs for teaching. Ellis (2017) maintains that teachers make decisions on how to practice FonF, and they should do so in a principled manner. This study's findings suggest that teachers make informed decisions when focusing on FSs compared to non-formulaic forms and use different pedagogical options in terms of the types of FonF, timing of FonF, etc., in providing incidental FonF to promote learners' (successful) uptake of FSs and non-formulaic forms targeted in FonF. Finally, this study was conducted in an EFL context with advanced adult EFL learners. Future studies could offer insights into FonF practices by examining FonF characteristics that predict successful uptake of FSs vs. non-formulaic forms in other settings with learners from different proficiency levels.

7. Acknowledgment

The author wishes to extend her gratitude to the teachers and learners who participated in this study.

8. References

American Council on the Teaching of Foreign Languages (2012). *ACTFL proficiency guidelines*. Retrieved from http://actflproficiencyguidelines2012.org/

Ayto, J. (2010). *Oxford dictionary of English idioms* (3rd ed.). Oxford, UK: Oxford University Press. doi: 10.1093/acref/9780199543793.001.0001

Benson, M., Benson, E., & Ilson, R. (2010). *The BBI combinatory dictionary of English: Your guide to collocations and grammar* (3rd ed.). Amsterdam, Philadelphia: John Benjamins. doi: https://doi.org/10.1075/z.bbi.07dic

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Brown, D. (2016). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*, *20*(4), 436-458. doi:10.1177/1362168814563200

Buerki, A. (2016). Formulaic sequences: A drop in the ocean of constructions or something more significant? *European Journal of English Studies*, *20*(1), 15–34. doi: https://doi.org/10.1080/13825577.2015.1136158

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.

Clare, A., & Wilson, J. J. (2016). *Speakout* (2nd ed.). UK: Pearson Education Limited.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. doi: https://doi.org/10.1177/001316446002000104

DeKeyser, R. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning*, *62*, 189–200. doi:10.1111/j.1467-9922.2012.00712.x

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206–225). Cambridge: Cambridge University Press. doi: https://doi.org/10.1017/CBO9781139524780.010

Egi, T. (2010). Uptake, modified output, and learner perceptions of recasts: Learner responses as language awareness. *The Modern Language Journal*, *94*, 1–21. doi:10.1111/j.1540-4781.2009.00980.x

Ellis, R. (2017). Oral corrective feedback in L2 classrooms. In H. Nassaji & E. Kartchava (Eds), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 3–18). Routledge. doi:10.4324/9781315621432-2

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, *51*(2), 281–318. doi: 10.1111/1467-9922.00156

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62. doi: https://doi.org/10.1515/text.1.2000.20.1.29

Gholami, L. (2021a). Incidental corrective feedback provision for formulaic vs. Non-formulaic errors: EFL teachers' beliefs and practices. *Language Awareness*, 1–32. doi:10.1080/09658416.2021.1943421

Gholami, L. (2021b). Oral corrective feedback and learner uptake in L2 classrooms: Non-formulaic vs. formulaic errors. *Language Teaching Research*, 136216882110215. doi:10.1177/13621688211021560

Gholami, L. (2021c). Incidental reactive focus on form in language classes: Learners' formulaic versus nonformulaic errors, their treatment, and effectiveness in communicative interactions. *Foreign Language Annals*, 1–26. doi: https://doi.org/10.1111/flan.12546

Gholami, L., & Gholami, J. (2018). Uptake in incidental focus-on-form episodes concerning formulaic language in advanced adult EFL classes. *Language Teaching Research*, *24*(2), 189–219. doi:10.1177/1362168818783442

Gholami, L., Karimi, M. N., & Atai, M. R. (2017). Formulaic focus-on-form episodes in adult EFL communicative interactions. *System*, *68*, 72–86. doi:10.1016/j.system.2017.06.015

Gurzynski-Weiss, L., & Baralt, M. (2015). Does type of modified output correspond to learner noticing of feedback? A closer look in face-to-face and computer-mediated task-based interaction. *Applied Psycholinguistics*, *36*, 1393–1420. doi: https://doi.org/10.1017/S0142716414000320

Hosmer, D. & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Howarth, P. (1998). The phraseology of learners' academic writing. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161–186). Oxford: Oxford University Press.

Hudson, J., & Wiktorsson, M. (2009). Formulaic language and the relater category-the case of *about*. In R. Corrigan, E. A. Moravcsik, H. Quali, & K. M. Wheatley (Eds.), *Formulaic language: Structure, distribution and historical change* (pp. 77–95). Philadelphia: John Benjamins.

Kecskes, I. (2016). Deliberate creativity and formulaic language use. In A. Capone, F. Lo Piparo, & M. Carapezza (Eds.), *Perspectives in pragmatics, philosophy & psychology* (pp. 3–20). doi:10.1007/978-3-319-43491-9_1

Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in System. *System*, *84*, 93–109. doi:10.1016/j.system.2019.05.006

Li, S., Zhu, Y. & Ellis, R. (2016). The effects of the timing of corrective feedback on the acquisition of a new linguistic structure. *The Modern Language Journal*, *100*(1), 276–295. doi:10.1111/modl.12315

Loewen, S. (2004). Uptake in incidental focus on form in meaning-focused ESL lessons. *Language Learning*, *54*(1), 153–188. doi: 10.1111/j.1467-9922.2004.00251.x

Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, *27*, 361–386. doi: https://doi.org/10.1017/S0272263105050163

Loewen, S. & Sato, M. (2019). Instructed second language acquisition and English language teaching: Theory, research, and pedagogy. In X. Gao (Ed.) *Second handbook of English language teaching* (pp. 1131–1148). Springer Nature Switzerland. doi:10.1007/978-3-319-58542-0_60-1.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie, & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press. doi: https://doi.org/10.1016/b978-012589042-7/50015-3

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition*, *20*, 51–81. doi: https://doi.org/10.1017/s027226319800103x

Lyster, R. (2001). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *51*, 265–301. doi:10.1111/j.1467-1770.2001.tb00019.x

Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics*, *32*, 111–129. doi: https://doi.org/10.1017/S0267190512000128

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, *32*(2), 129–148. doi: https://doi.org/10.1093/applin/amq035

Myles, F., & Cordier, C. (2017). Formulaic sequence (FS) cannot be an umbrella term in SLA. *Studies in Second Language Acquisition*, *39*(1), 3–28. doi:10.1017/s027226311600036x

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam: Benjamins. doi: 10.1075/scl.14

Panova, I., & Lyster, R. (2002). Patterns of Corrective Feedback and Uptake in an Adult ESL Classroom. *TESOL Quarterly*, *36*(4), 573-595. doi:10.2307/3588241

Quinn, P. G., & Nakata, T. (2017). The timing of oral corrective feedback. In H. Nassaji & E. Kartchava (Eds), *Corrective feedback in second language teaching and learning*, (pp. 35–47). Oxon, UK: Routledge. doi:10.4324/9781315621432-4

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu, HI: University of Hawaii, Second Language Teaching & Curriculum Center.

Siefring, J. (2005). *The Oxford dictionary of Idioms* (2nd ed.). Oxford: Oxford University Press.

Simpson-Vlach, R., & Ellis, C. N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512. doi: 10.1093/applin/amp058

Sinclair, J. M. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Siyanova-Chanturia, A. (2019). Formulaic language. In A, Siyanova-Chanturia, & A. Pellicer-Sanchez (Eds), *Understanding formulaic language: A second language acquisition perspective* (pp. 1–15). New York, NY: Routledge. doi:10.4324/9781315206615-1

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), Principle and practice in applied linguistics: Studies in honor of H.G. Widdowson (pp. 125–144). Oxford: Oxford University Press.

Warren, B. (2005). A model of idiomaticity. *Nordic Journal of English Studies*, *4*, 35–54.

Wood, D. (2020). Classifying and identifying formulaic language. In S. Webb (Ed.), *Routledge handbook of vocabulary studies* (pp. 30 – 45). Routledge: London.

Wray, A. (2019). Concluding question: Why don't second language learners more proactively target formulaic sequences. In A, Siyanova-Chanturia, & A. Pellicer-Sanchez (Eds), *Understanding formulaic language: A second language acquisition perspective* (pp. 19–37). New York, NY: Routledge. doi: https://doi.org/10.4324/9781315206615-1

Wray, A., & Namba, K. (2003). Use of formulaic language by a Japanese-English bilingual child: A practical approach to data analysis. *Japan Journal of Multilingualism and Multiculturalism*, *9*(1), 24–51.

Wulff, S. (2019). Acquisition of formulaic language from a usage-based perspective. In A, Siyanova-Chanturia, & A. Pellicer-Sanchez (Eds), *Understanding formulaic language: A second language acquisition perspective* (pp. 19–37). New York, NY: Routledge. doi:10.4324/9781315206615-2

Xu, J. (2015). *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech* (Doctoral dissertation).Retrieved from http: //lib.dr.iastate.edu/etd/14875/.

Zhao, S., & Bitchener, J. (2007). Incidental focus on form in teacher-learner and learner-learner interactions. *System*, *35*(4), 431–447.doi: http://dx.doi.org/10.1016/j.system.2007.04.004

Appendices

Appendix A: Criteria for the identification of FSs (adopted from Wray & Namba, 2003: 29–33).

Criterion	
Example	$\kappa = .80$
1. By my judgment, there is something grammatically unusual about this wordstring.	*rains cats and dogs*, "*rain*" is intransitive
2. By my judgment, part or all of the wordstring lacks semantic transparency.	*curry favor* is non-compositional
3. By my judgment, this wordstring is associated with a specific situation and/or register.	*I wonder if would mind ...*
4. By my judgment, the wordstring as a whole performs a function situation and/or register	*on the other hand*
5. By my judgment, the speaker/writer has accompanied this of wordstring with an action, use of punctuation, or phonologica pattern that gives it special status as a unit, and/or is repeating something s/he has just heard or read.	Idiomatic reading of *pick-you-own vegetables*
6. By my judgment, the speaker has marked this wordstring grammatically or lexically in a way that gives it special status as a unit.	I have just learned *pin money*
7. By my judgment, there is a greater than chance-level probability that the speaker will have encountered this precise formulation before.	*false teeth*
8. By my judgment, although this wordstring is novel, it is a clear derivation, deliberate or otherwise, of something that can be demonstrated to be formulaic in its own right.	*I slept like a twig* instead of *I slept like a log*
9. By my judgment, this wordstring is formulaic, but it has been unintentionally applied inappropriately	*all's well that ended well*
10. By my judgment, this wordstring contains linguistic material that is too sophisticated, or not sophisticated enough, to match the speaker's general grammatical and lexical competence.	*outstay one's welcome*
11. By my judgment, there is an underlying frame and one or more gaps in this wordstring. The frame is formulaic and the gaps can be filled with any lexical items.	NP *feel*-TENSE ... *to* + infinitive

Note. κ refers to the value of inter-coder reliability.

Linguistic focus Description Example κ = .89

Appendix B: Linguistic foci of formulaic FFEs (adapted from Xu, 2015: 86).

Linguistic focus	Description	Example
Example		κ = .80
A. The entire FS	Overall form, meaning, usage of an FS, or both lexical selection and lexical formation	S: ؟دهکده جهانی /dehkæ-deh dʒæhʌni:/ T: *global village*
B. Lexical selection	The meaningfulness of combination of content words in an FS.	S: *became daggers at me* T: *looked daggers*
1. Unacceptable selection	Deviation in the selection of (a) content word(s).	S: *make a chance* T: *take a chance*
2. Substandard selection	Preference of a content word over possible others	S: *It's very functional* T: *yeah, fully functional*
C. Lexical formation	Morphological elements in an FS	
1. Word form	Inflectional and/or derivational deviation involving substitution, omission, or insertion of singular, plural, restricted tense or morpho-syntactic elements, etc	S: *I said "stick to your gun"* T: *stick to your GUNS*
2. Function word	Deviation involving substitution, omission, or insertion of a preposition, a particle, etc.	S: *to hang off with friends* T: *hang OUT*
3. Determiner	Deviation involving substitution, omission, or insertion of a determiner.	S: *add insult to an injury* T: *add insult to injury*

Note. κ refers to the value of inter-coder reliability.

Appendix C: Syntactic patterns of lexical collocations (adopted from Xu, 2015: 80).

Type	Examples	κ = .86
1. Adjective and noun (ADJ-N)	S: *I hate crowded traffic* T: *heavy traffic*	
2. Adverb and adjective (ADV-ADJ)	S: *It was clearly clear that* T: *perfectly clear*	
3. Adverb and verb (ADV-V)	S: *I hugely advise using this* T: *strongly advise*	
4. Noun-noun (N-N)	S: *I want to know the first cause of it* T: *we say root cause*	
5. Noun of noun (N-of-N)	S: *a group of wolves* T: *a pack of wolves*	
6. Noun and verb (N-V)	S: *and lions barked* T: *lions roar not bark!*	
7. Verb and noun (V-N)	S: *and, I made photograph* T: *made or took?*	
8. Phrasal verb and adverb (PHV-ADV)	S: *I had to wake early* T: *wake up, right?*	
9. Noun and phrasal verb (N-PHV)	S: *when sunflowers come* T: *we say flowers come out*	
10. Phrasal verb and noun (PHV-N)	S: *we have to adhere by rules* T: *adhere to rules*	

Note. κ refers to the value of inter-coder reliability.

Appendix D: Patterns of grammatical collocations (adopted from Benson et al., 2010: XIX–XXX).

Type	Example	κ = .88
1. Noun and preposition (N-Prep)	S: *comply by guidelines* T: *comply with*	
2. Noun and to-infinitive (N-to-Inf)	S: *I felt a compulsion ding it* T: *a compulsion to do*	
3. Noun and that-clause (N-that-clause)	S: *I took an oath about abiding by the regulations* T: *took an oath that you would abide by the regulations*	
4. Preposition and noun (Prep-N)	S: *I broke it based on accident* T: *by accident*	
5. Adjective and preposition (Adj-Prep)	S: *the plants are native in north of Iran* T: *native to north of Iran*	
6. Adjective and to-infinitive (Adj-to-Inf)	S: *the accident was bound for happening* T: *bound to happen*	
7. Adjective and that-clause (Adj-that-clause)	T: *I was afraid to fail the exam* S: *I was afraid that I would fail the exam*	
8. Verb and preposition (V-Prep)	S: *apologize from my sister* T: *apologize to*	

Note. κ refers to the value of inter-coder reliability.

Appendix E: Taxonomy of lexical bundles (adopted from Simpson-Vlach & Ellis, 2010: 37–42).

Pragmatic function	AFL	Lexical bundle	*FTW* rank	κ = .92
A. Referential Expressions	Core	*as a function of*	1.19	
1. Specification of attributes	Core	*with respect to*	1.26	
a) Intangible framing attributes	Written	*in accordance with*	1.36	
2. Contrast and comparison	Written	*on the other hand*	2.84	
	Core	*as opposed to*	1.02	
3. Deictics and locatives	Spoken	*at this point*	0.61	
4. Vagueness markers	Spoken	*and so forth*	0.80	
B. Stance Expressions				
1. Hedges	Spoken	*in a sense*	0.56	
2. Epistemic stance	Core	*according to*	0.18	
C. Discourse Organizing Functions				
1. Metadiscourse and textual reference	Written	*at the outset*	0.51	
2. Topic elaboration				
a) Topic elaboration: cause & effect	Written	as a consequence	0.50	
3. Discourse markers	Core	*in other words*	1.90	
	Spoken	*by the way*	0.45	

Notes. *FTW* stands for formula teaching worth. κ refers to the value of the inter-coder reliability.

Appendix F: Patterns of compounds (adopted from Wood, 2015: 47).

Type	Description	Example $\kappa = .95$
A. Closed form	The two components are written as one.	*brainstorm*
B. Hyphenated form	The two components are separated by hyphens.to	*water-proof*
C. Open form	The two components are written separately.	*Bullet point*

Note. κ refers to the value of inter-coder reliability.

Gender pairings in EFL child task-based interaction

María Martínez-Adrián
Department of English, German and Translation and Interpretation Studies
University of the Basque Country UPV/EHU
maria.martineza@ehu.eus

María Juncal Gutiérrez-Mangado
Department of English, German and Translation and Interpretation Studies
University of the Basque Country UPV/EHU
junkal.gutierrez@ehu.eus

Abstract

Few studies have tackled gender differences in second language (L2) interaction, and particularly, the effect of gender pairings on learning opportunities operationalized as Language Related Episodes (LREs) has been scarcely looked into (see Azkarai, 2015b; Azkarai & García-Mayo, 2012; Ross-Feldman, 2005, 2007). Additionally, these studies have targeted adult L2 learners and to our knowledge, no studies so far have been conducted with children. This paper will try to fill these gaps by analysing the effect of gender pairings on the occurrence, nature and resolution of LREs in a storytelling task performed by 10-12-year-old children. More specifically, it explores whether there are any differences between same-gender and gender-mismatched dyads, and between same-gender dyads (male-male vs female-female). Results show that type of pairing affects LRE production and resolution as more LREs were initiated and resolved in matched-gender dyads than in mixed-gender dyads. However, gender-pairing did not influence the outcome of the resolution, as a still low rate of targetlikeness was obtained in mixed- and matched-gender dyads. Additionally, no statistically significant differences were obtained between same-gender dyads, but female-female dyads were found to be more concerned with getting the message across and oriented to higher accuracy in the resolutions of meaning-related episodes.

Keywords: gender pairings; young learners; language related episodes; task-based interaction; English as a foreign language.

Resumen

Pocos han sido los estudios que han abordado las diferencias de género durante la interacción en una segunda lengua (L2), y particularmente, el efecto de los emparejamientos según el género de los aprendices en la producción de Episodios Relacionados con el Lenguaje (ERLs) ha sido poco investigado (véase Azkarai, 2015b; Azkarai y García Mayo, 2012; Ross-Feldman, 2005, 2007). Del mismo modo, dichas investigaciones han examinado aprendices adultos, y no parecen existir investigaciones con niños. Este trabajo analiza el efecto de los emparejamientos según el género de sus miembros en la ocurrencia, naturaleza y resolución de ERLs surgidos durante la narración oral de una historia por parte de niños de 10 a 12 años. Concretamente, explora si existen diferencias entre parejas de mismo y distinto género, y entre parejas de mismo género (niño-niño vs. niña-niña). Los resultados demuestran que el tipo de emparejamiento afecta a la producción y resolución de ERLs ya que un mayor número de ERLs fueron iniciados y resueltos en parejas del mismo género frente a parejas de distinto género. Sin embargo, el tipo de emparejamiento (mismo vs. distinto género) no tuvo repercusión a nivel de episodios resueltos de forma correcta. Por otro lado, la comparación entre parejas del mismo género no conllevó diferencias significativas, aunque el análisis descriptivo sí indicó que las parejas formadas únicamente por chicas parecían estar más implicadas en transmitir el mensaje y más orientadas hacia una resolución correcta de los episodios relacionados con el léxico.

Palabras clave: emparejamientos por género; jóvenes aprendices; episodios relacionados con el lenguaje; interacción durante la realización de tareas; inglés como lengua extranjera.

1. Introduction

The study of the effect of gender on second language (L2) performance has been approached from two different social perspectives. On the one hand, studies examining gender attitudes and learning outcomes have claimed that language skills in the case of females develop quicker than in males. In addition, females show a higher inclination to study foreign languages and have more positive attitudes and greater motivation towards languages (Pavlenko & Piller, 2008; Spolsky, 1989; Sunderland, 2000). On the other hand, another strand of studies within the Interactionist framework, though more limited, has explored the effect of dyad type on task-based interaction. In this respect, mixed results in terms of amount of negotiation have been obtained in English as a second language (ESL) and English as foreign language (EFL) settings (i.e. Azkarai & Imaz Agirre, 2017; Oliver, 2002; Shehadeh, 1994, 1999). Also within the Interactionist framework, other studies have explored the impact of collaborative

tasks on attention to form operationalized as Language Related Episodes (LREs). This construct has been defined as any part of the dialogue in which students talk about the language they are producing, question their language use, or other- or self-correct' (Swain, 1998: 70). However, to the knowledge of the authors, this line of research has received scant attention in the literature and no studies have been conducted with young learners. Studies carried out with adult learners on LRE production have reported the inexistence of differences between mixed and matched-gender pairings except for resolution. This line of research is fully justified in the case of young learners if we consider their different engagement in the language learning process from adult learners (Oliver & Azkarai, 2017). The analysis of the impact of gender on the production of LREs during child task-based interaction may shed more light on the most efficient learner pairings and may contribute to maximize the learning potential of collaborative tasks for attention to form. Exploiting the resources at hand seems to be even more important in EFL contexts, also considered low input settings (Pinter, 2011) where learners cannot benefit from the large amounts of exposure offered in ESL settings.

The paper is organized as follows. Section 2 provides an overview of empirical findings related to the effect of gender on interaction as well as on language learning opportunities (LREs). The study is described in section 3. Results are subsequently shown and discussed in sections 4 and 5, respectively. Section 6 concludes the paper.

2. Literature review

2.1 Gender and interaction

The difference in communicative patterns observed in males and females during first language (L1) interaction has its origin in the socialization process during childhood (Azkarai, 2013). Children spend large amounts of time playing and interacting with same-gender friends (Maltz & Borker, 1982; Tannen, 1994). This entails that boys and girls have different experiences and learn many skills, competencies, and interests in their interactions with same-sex peers (Hanish & Fabes, 2014).

Sex segregation emerges early in life and by preschool, about half of children's interactions are with same-sex peers (Fabes, Martin, & Hanish, 2003), and it persists through preadolescence and adulthood (Mehta & Strough, 2009, 2010). It has important consequences for the development of communication skills. In particular, males and females develop different ways to express themselves (Maltz & Borker, 1982). For example, research has shown how females in matched-gender dyads tend

to talk more overall and discuss fewer topics than males who discuss many topics briefly. Additionally, males' discussions are more abstract than females' (Tannen, 1990). In mixed-gender dyads, males have been reported to discuss a wider range of topics than in male-only groups (Aries, 1976), and have also been found to have more opportunities to participate and control conversations than females (see Bohn & Stutman, 1983; West & García, 1988; West & Zimmerman, 1983; Zimmerman & West, 1975). These behaviours observed in males and females might have an impact on L2 learning opportunities during task-based interaction.

In L2 acquisition, gender has received limited attention within the interactional approach (Long, 1996) and mixed findings have been obtained, which could be ascribed to the interaction between cultural norms and gender. As discussed by Shehadeh (1999), it is more acceptable in some cultures and subcultures than in others for men and women to communicate freely and casually with each other at work and in social situations (p. 260). As a result, the different cultures of talk imposed on males and females that could vary across contexts (see Pavlenko, 2001 in this respect) should not be overlooked and research should place gender differences within specific contexts (Gu, 2013). Studies conducted in ESL and EFL contexts with adults and children have specifically tested the effect of dyad-type on task-based interaction. Some of these studies have yielded differences between mixed and same-gender dyads, while other investigations have not found differences between dyad-types. In the case of studies conducted with adults in ESL contexts, Shehadeh (1994) in an investigation with L2 learners with a variety of L1s (i.e. Afrikaans, Arabic, Bengali, Cantonese, Farsi, French, Greek, Italian, Korean, Mandarin, Portuguese, Serbo-Croat and Spanish) reported the existence of more opportunities for males to produce comprehensible output and to request clarifications in mixed-gender dyads, while females appeared to have more opportunities to produce comprehensible output in matched-gender dyads. In Gass and Varonis (1986) more negotiations in the case of L1 Japanese learners of L2 English occurred in mixed-gender dyads than in matched-gender dyads. Other studies conducted with L1 Japanese adult learners of L2 English such as Pica et al. (1991) have not found significant differences between types of dyads. As for children, to our knowledge, only Oliver (2002) has focused on gender and interaction in ESL settings, where she reported non-significant differences between male-male and female-female dyads.

In EFL settings, particularly in the Spanish context in which the present study was conducted, more research also exists with adults than with children but the findings have also been mixed. Adult females negotiate more in matched-gender dyads than in mixed-gender settings (Alcón, 1994) and tend to resort to their L1 more than males (Azkarai, 2015a). In mixed-gender dyads, males have been reported to interrupt and talk more frequently and introduce a greater number and wider range of topics (Alcón,

1994). In other studies with adults, such as Alcón and Codina (1996), overall gender did not affect negotiation but some task effects were found when comparing female-female dyads and mixed-gender dyads, as also reported in Birjandi and Tabatabaei (2010). In the information-gap task administered, females negotiated more than males, while no differences existed in the discussion task they performed. As in the ESL context, little research has also been conducted with children in EFL settings along these lines (Azkarai & Imaz Agirre, 2017) and a call has been made by several authors for research with children as it could have implications aimed at maximizing contact hours with learners. Knowing about the impact of gender on interaction might help teachers to organize activities that take advantage of the most convenient pairings. In Azkarai and Imaz Agirre (2017) with 3rd and 4th year primary-school learners, younger learners were found to negotiate more in same-gender dyads, while older learners did so in mixed-gender dyads.

2.2 Gender and LREs

Other studies within the interactional framework have specifically tested the effect of gender on the production of LREs, but are more limited in number (Azkarai, 2015b; Azkarai & García Mayo, 2012; Ross-Feldman, 2005, 2007). As stated by Ross-Feldman (2007: 57), any effect of gender on the incidence and resolution of LREs might differentially influence the language learning opportunities available to male and female learners as a result of engaging in interaction. Ross-Feldman (2005, 2007) examined the incidence and resolution of LREs of 32 male and 32 female ESL learners while working in mixed and matched-gender dyads in three different tasks (picture differences, picture placement and picture story task). Each participant interacted in both mixed and matched-gender dyads. The analysis consisted in the comparison of different dyad types, comparison of males to females in mixed-gender dyads and the comparison of the interaction of the same learner in matched and mixed-gender dyads. The results showed that as for incidence, whether participants interacted in mixed or matched-gender dyads did not significantly influence the production of LREs. However, in terms of resolution, in mixed-gender dyads, LREs initiated by males were resolved more often than LREs initiated by females. Similarly, when comparing the interaction of the same learner in mixed- and matched-gender dyads, LREs initiated by males were resolved more often in mixed-gender dyads than in matched-gender dyads, and LREs initiated by females were resolved more often in matched-gender than in mixed-gender dyads. In the light of the results shown in this study, Ross-Feldman concludes that males have more opportunities to learn the target language (TL) when working in mixed-gender dyads, while females have more opportunities in matched-gender dyads, results which have also been supported by Aries (1976) and Holmes (1994) and also by research in education (Sadker & Sadker, 1994).

More recently, Azkarai & García Mayo (2012), and Azkarai (2015b) tested the effect of dyad-type on the production of LREs by EFL learners along the same lines as Ross-Feldman (2005, 2007). In particular, Azkarai (2015b) which included a greater number of participants than Azkarai & García Mayo (2012), examined 22 males and 22 females while completing four different tasks. They worked first in matched-gender dyads and then in mixed-gender dyads. When comparing males and females in terms of number, nature (meaning and form), and outcome of LREs, the analysis of the results did not yield statistically significant differences. These findings support Ross-Feldman (2005, 2007) and Azkarai & García Mayo (2012) when examining incidence and nature of LREs. However, they contrast with Ross-Feldman (2005, 2007) in terms of outcome, as in her study LREs initiated by males were resolved significantly more often than those initiated by females. The different social context in Ross-Feldman (2005, 2007) and Azkarai (2015b) might be the underlying reason for these differences. Namely, socialization patterns in South America, where the study by Ross-Feldman (2005, 2007) was conducted, are different from the ones in Europe, where the investigations by Azkarai & García Mayo (2012) and Azkarai (2015b) were carried out. Azkarai (2015b) also focused on the possible differences in LREs initiated by males and females depending on the gender of the interlocutor. No differences were found in nature and occurrence of LREs when females worked in mixed or matched-gender dyads, but differences emerged in the case of outcome, as females resolved more LREs when working with males. Similarly, no differences in incidence and nature of LREs were obtained when males worked either in mixed and matched-gender pairings. However, males ignored more LREs in matched-gender dyads. These results also contrast with Ross-Feldman (2005, 2007), as in her study working in mixed dyads seemed to favour male learners while working in matched-gender dyads seemed to be more profitable for females. In conclusion, males and females focused similarly on meaning and form in same-gender and mixed-gender dyads. But in terms of resolution, males and females seem to benefit more when working in mixed-gender pairings.

As the dearth of studies described above suggests, the impact of type of pairing on the production, nature and resolution of LREs has been scarcely looked into. In particular, to our knowledge, just four studies have explored males' and females' behaviour working in mixed and matched gender dyads (see Azkarai, 2015b; Azkarai & García-Mayo, 2012; Ross-Feldman, 2005, 2007). In addition, all these studies have targeted adult learners, and no investigations along these lines have been conducted with children. In this respect, a call has been made for more research on young L2 learners given the specificity of their language learning approach and their peer exchanges in this population (Muñoz, 2007; as cited in Plavadevall Ballester & Vraciu, 2020). This article will try to fill these gaps by analyzing the oral interaction of young Spanish EFL learners on the basis of their gender and the language learning

opportunities (operationalized as LREs) available to them during task-based interaction. In particular, this paper aims to study gender differences in the occurrence, nature and outcome of LREs on the basis of Azkarai (2015b). More specifically, whether there are any differences between same-gender and gender-mismatched dyads, and between same-gender dyads (male-male vs female-female).

3. The Study

The present study is part of a bigger project on task-modality effects among young learners. In particular, this study investigates the effect of gender pairings on the production of LREs by 10-12 year old children in a primary school from the Basque Country (Spain) with the aim of answering the following research questions:

(1) Are there any differences in terms of number, types and outcome of LREs among different types of gender-pairings?

1.a. Are there any differences between same-gender and mixed-gender dyads?

1.b. Are there any differences between male-male and female-female dyads?

(2) What are the more common types of LREs (nature and outcome) in each dyad group?

3.1 Participants

The participants in the study were 56 (19 females, 37 males) child EFL learners, aged 10-12, from two intact classes[1] in the 5th year (14 males, 8 females) and three in the 6th year (23 males, 11 females) of primary education in the Basque Autonomous Community in Spain, a European country offering balanced opportunities for males and females to interact (Azkarai, 2015b). The participants were attending the educational model D, by which the vast majority of instruction takes place in Basque (the minority language of the area in which these learners live). Spanish Language and its Literature is taught in Spanish, and English as a language subject and content subjects such as science, arts and crafts or physical education are taught in English. This model was originally created as a language maintenance programme for native speakers of Basque but nowadays, this programme includes a high number of L1

1 In this project we decided to preserve 'ecological validity' (Van Lier, 1988) by gathering data from intact groups instead of a tight control of extraneous variables as in true experiments. In this respect, this study could be best described as exploratory. The findings obtained will be concurrently validated in further investigations with a more balanced sample.

Spanish students who can achieve a level of functional competence in Basque alongside the 'usual' level of knowledge of the L1 (Zalbide & Cenoz, 2008: 9). Thus, participants in this model are considered Basque-Spanish bilinguals who learn their third language (L3) chronologically speaking, as a foreign language (Cenoz, 2009).

They started learning L3 English as a school subject at around age 4 and since Grade 3 they have been enrolled in a CLIL programme in which English is used as a language of instruction for the aforementioned content subjects. At the time of data collection, they were receiving 3 hours of EFL instruction and 2 to 4 hours a week of CLIL instruction.

The fifth grade children had accumulated approximately 777 hours of instruction in English, and the sixth grade children 962 hours. Before data collection, the children were administered a linguistic background questionnaire and an English placement test (KET) whose results indicated that participants could be considered beginners (A1-A2)[2]. Participant main characteristics together with their proficiency scores are shown in Table 1.

Table 1. Participants' characteristics

	LH5		LH6	
	Females	Males	Females	Males
Mean Age at testing	9.5 (SD: 0.53)	9.92 (SD: 0.26)	11 (SD: 0.77)	10.92 (SD: 0.57)
Mean AoA English	4.12 (SD: 0.83)	4.07 (SD: 1.68)	4.81 (SD: 1.16)	4.88 (SD: 1.78)
Test score (Max. score in the test=85)	24.62 (SD: 6.45)	27.21 (SD: 11.44)	46.63 (SD: 14.41)	42.96 (SD: 18.20)

Note: SD= standard deviation; AoA= age of acquisition.

The children were paired into 28 dyads to perform the collaborative tasks administered in the project. Even if all the participants in the project were considered beginners as their scores fell in the range A1-A2, they were paired with children from the same grade and matched in same-proficiency dyads. This yielded the following gender-pairings: 7 dyads were gender mismatched (2 from the 5th year and 5 form the 6th year); 6 were female-female (FF) dyads (3 from the 5th year and 3 from the 6th year) and 15 were male-male (MM) dyads (6 from the 5th year and 9 from the 6th year). As another safeguard to control for proficiency measures, a Mann-Whitney U Test was computed

2 Basic users according to the Common European Framework of Reference for languages http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf

to compare male and female participants which revealed the inexistence of statistically significant differences in terms of proficiency (Z=-0.666; p-value=0.506). Thus, any potential gender differences during task-based interaction cannot be ascribed to their distinct proficiency level.

3.2. Materials and procedure

The task described in the present paper is part of a wider project aimed at investigating collaborative tasks (speaking and speaking+writing tasks) in primary-school children. In particular, this project is making a valuable contribution to the still limited study of young learners in foreign language contexts by investigating task-modality effects attested in previous research with adults (see García Mayo & Azkarai, 2016)[3]. Children's uniqueness in their engagement in the learning process clearly reinforces the specificity of L2 research with this population (Mackey & Gass, 2005; as cited in Oliver & Azkarai, 2017). Thus, in this project similar tasks to the ones employed with adults have been designed so as to investigate whether similar results would be obtained and whether any potential adaptations should be included so as to maximize children's opportunities for learning (García Mayo, 2017).

For the task reported here (speaking task), each pair was video-taped while ordering a set of six pictures (see Appendix 1) before orally narrating the events depicted in them in story mode. This story was chosen from an appropriate level English book, Sparks 1 (House & Scott, 2009), with the teachers' approval and it was pilot-tested with similar age children prior to its administration.

Tasks where pictures have to be ordered before narrating the story depicted have been widely used in ESL and EFL research (García Mayo & Lázaro Ibarrola, 2015; García Mayo & Hidalgo Gordo, 2017; Lázaro Ibarrola & Azpilicueta Martínez, 2015; Martínez-Adrián, 2020; Pica, Kang, & Sauro, 2006; Pica et al. 1993) because they create real opportunities for students to interact with each other in order to agree on the order of the pictures and the possible story which they depict. This type of oral tasks are considered more immediate tasks, with very little planning and editing time, and their primary focus is meaning (Payant & Kim, 2019). In contrast, tasks that combine speaking and writing modalities seem to demand higher levels of accuracy because of the extra processing time learners have to reflect on their production

3 In the first set of investigations conducted within the project (Martínez-Adrián & Gallardo-del-Puerto, forthcoming; Gallardo-del-Puerto & Martínez-Adrián, in press; Gutiérrez-Mangado & Basterrechea, forthcoming) and so as to obtain an overall picture of task-modality effects, we have compared different tasks and modalities, leaving the interface between modality and other grouping variables for further research.

(i.e. Adams & Ross-Feldman, 2008; Azkarai & García Mayo, 2015; García Mayo & Azkarai, 2016; Niu, 2009; Payant & Kim, 2019). However, these studies that have compared speaking tasks to speaking+writing tasks have not controlled for the level of accuracy that both tasks demand. Thus, in order to solve the limitation of previous research and taking into account that the framing of the task could overrule the inherent focus of the task (Philp, Walter, & Basturkmen, 2010), in the present project, instructions were kept constant by asking learners to attend to accuracy in the tasks administered. More specifically, participants performing the speaking task were asked to pay special attention to the way they told the story in that it had to make sense but also be linguistically accurate.

The task was carried out in a quiet room with the presence of the investigator, who explained the task and gave precise instructions in the TL English as to how to carry it out. The children were asked to act naturally and were also informed that they could seek out each other's help (but not the assistant's) in case of difficulties with vocabulary.

3.3. Data codification and analysis

Participants' oral interactions were transcribed verbatim and coded in CHILDES (MacWhinney, 2000) format. The transcriptions were analysed in order to identify all LREs. LREs were classified on the basis of their nature (meaning vs. form) and outcome (resolved vs. unresolved). Resolved LREs were further categorized as target-like or non-target-like (Azkarai, 2015b; Ross-Feldman, 2005, 2007). They were first coded by one of the researchers and then checked by another one. Mismatches between both researchers were resolved through discussion.

LREs were classified as meaning-related when the interaction involved the meaning or use of a word and form-related when they related to the grammatical form (phonology, morphology, syntax, spelling or preposition) of a word or sentence. LREs were further coded as resolved when a solution was provided to the issue raised in the LRE or unresolved when proceeded with the task without finding a solution (1). Solved LREs were further classified as target-like, when the solution suggested by the interlocutor's feedback was correct (2) or as non-target-like when the proposed solution was incorrect (3).

(1) Meaning focused; unresolved

1 HL eeeh, I have for ... my sister and ... af eh ... ¿cómo se dice coser? [how do you say to sew?]

2 AR I don't know.

(2) Form focused; solved; target-like

1 AC eh... once upon a time eh ... eh ... there is two friends and she eh and .... and.

2 MG they are.

(3) Meaning focused; solved; non-target-like

1 SG she's making a .... Pan, panpin [doll (in Basque)], pumpkin?

2 ER no, eso es calabaza [no, that means pumpkin] laughter. ¿Cómo se dice? [how do you say it?]

3 SG muñeco [doll]... and the ....panpin, yo qué sé! [doll ... what do I know!]

4 ER then, panpin, eh?

As for the statistical analysis of the data, both descriptive and inferential statistical procedures were conducted. In the case of descriptive analyses, the number of LREs, the number of LRE types as well as their percentages, mean scores, and standard deviations were calculated. Regarding inferential analyses, since the data did not follow a normal distribution for most of the variables examined, non-parametric tests were computed. Between-group comparisons were analysed using the Kruskal Wallis and Mann-Whitney U tests, while Friedman and Wilcoxon Signed Rank tests were used for within-group comparisons. Statistical probability values were marked at below 0.05 for significant differences and below.09 for marginal differences[4].

4. Results

In this section we will show the results of the analyses computed to answer the two research questions[5]. First, the analyses for the intergroup comparisons (research question 1) will be shown. Table 2 illustrates the results for incidence and nature of LREs:

4 Given the lack of studies targeting young learners along these lines, marginal differences could potentially uncover tendencies that could be further examined in a larger sample.

5 Note that as in other studies conducted within the very same project, 5th and 6th year learners have been examined jointly as their proficiency scores were in the range A1-A2 (beginner level). This decision aligns with previous research targeting young learners in which proficiency did not play a major role when comparing 5th and 6th year learners due to the small gap in terms of proficiency between the years tested (i.e. Gallardo-del-Puerto, Basterrechea, & Martínez-Adrián, 2020). If a wider gap in proficiency had been attested between grades (as suggested by these previous investigations), we would have explored and discussed the effect of different proficiency levels.

Table 2. Number, percentages, means and standard deviations of LREs initiated by same- and mixed-gender dyads

	Same-gender dyads	Mixed-gender dyads
Incidence	Number 98 Mean 4.67 Standard Deviation 3.10	Number 11 Mean 1.57 Standard Deviation 1.51
Nature of LREs		
Meaning-related	Number 91 Percentage 92.86% Mean 4.33 Standard Deviation 2.90	Number 11 Percentage 100% Mean 1.57 Standard Deviation 1.51
Form-related	Number 7 Percentage 7.14% Mean 0.33 Standard Deviation 0.79	Number 0 Percentage 0 Mean 0 Standard Deviation 0

The results (Table 2) revealed that same-gender dyads produced more LREs than mixed-gender dyads (Z=-2.324, p=0.020). As for nature, most LREs were meaning-related in both types of dyads but same-gender dyads significantly produced more (Z=-2.244, p=0.025).

Details of the outcome of the LREs are shown in Table 3:

Table 3. Number, percentages, means and standard deviations of outcomes of LREs initiated by same- and mixed-gender dyads

	Same-gender dyads	Mixed-gender dyads
Meaning-related		
Resolved	Number 71 Percentage 78.02% Mean 3.38 Standard Deviation 2.57	Number 7 Percentage 63,64% Mean 1.00 Standard Deviation 1.41
Target-like	Number 33 Percentage 46.48% Mean 1.57 Standard Deviation 1.59	Number 3 Percentage 42.86% Mean 0.43 Standard Deviation 0.78
Non-target-like	Number 38 Percentage 53.52% Mean 1.81 Standard Deviation 1.99	Number 4 Percentage 57.14% Mean 0.57 Standard Deviation 0.78
Unresolved	Number 20 Percentage 21.98% Mean 0.95 Standard Deviation 1.07	Number 4 Percentage 36.36% Mean 0.57 Standard Deviation 0.97
Form-related		
Resolved	Number 7 Percentage 100% Mean 0.33 Standard Deviation 0.79	Number 0 Percentage 0 Mean 0 Standard Deviation 0
Target-like	Number 5 Percentage 71.43% Mean 0.24 Standard Deviation 0.7	Number 0 Percentage 0 Mean 0 Standard Deviation 0
Non-target-like	Number 2 Percentage 28.57% Mean 0.10 Standard Deviation 0.43	Number 0 Percentage 0 Mean 0 Standard Deviation 0
Unresolved	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0

The results showed that same-gender dyads resolved meaning-related LREs more frequently ($Z=-2.261$, $p=0.024$) and that their resolution was more often target-like ($Z=-2.051$, $p=0.040$) than mixed-gender dyads. With respect to LREs resolved in a non-target-like manner and the number of unresolved meaning LREs, no differences were found. As regards form-related LREs, no differences were found between the two groups.

A second set of analyses to answer the first research question was carried out comparing MM, FF and mixed-gender dyads as regards incidence, nature and outcome of LREs (Table 4).

Table 4. Number, percentages, means and standard deviations of number, type and outcomes of LREs initiated by MM, FF and mixed-gender dyads

	Same-gender dyads		Mixed-gender dyads
	MM	FF	
Nº of LREs	Number 66 Mean 4.40 Standard Deviation 3.13	Number 32 Mean 5.33 Standard Deviation 3.20	Number 11 Mean1.57 Standard Deviation 1.51
Meaning-related	Number 59 Percentage 89.39% Mean 3.93 Standard Deviation 2.78	Number 32 Percentage 100% Mean 5.33 Standard Deviation 3.20	Number 11 Percentage 100% Mean 1.57 Standard Deviation 1.51
Resolved	Number 46 Percentage 77.97% Mean 3.07 Standard Deviation 2.52	Number 25 Percentage 78.13% Mean 4.17 Standard Deviation 2.78	Number 7 Percentage 63.64% Mean 1 Standard Deviation 1.41
Target-like	Number 18 Percentage 39.13% Mean 1.20 Standard Deviation 1.26	Number 15 Percentage 60% Mean 2.50 Standard Deviation 2.07	Number 3 Percentage 42.86% Mean 0.43 Standard Deviation 0.78
Non-target-like	Number 28 Percentage 60.87% Mean 1.87 Standard Deviation 2.03	Number 10 Percentage 40% Mean 1.67 Standard Deviation 2.06	Number 4 Percentage 57.14% Mean 0.57 Standard Deviation 0.78
Unresolved	Number 13 Percentage 22.03% Mean 0.87 Standard Deviation 1.18	Number 7 Percentage 21.87% Mean 1.17 Standard Deviation 0.75	Number 4 Percentage 36.36% Mean 0.57 Standard Deviation 0.97
Form-related	Number 7 Percentage 10.61% Mean 0.47 Standard Deviation 0.91	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0
Resolved	Number 7 Percentage 100% Mean 0.47 Standard Deviation 0.91	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0

Target-like	Number 5 Percentage 71.43% Mean 0.33 Standard Deviation 0.81	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0
Non-target-like	Number 2 Percentage 28.57% Mean 0.13 Standard Deviation 0.52	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0
Unresolved	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0	Number 0 Percentage 0 Mean 0 Standard Deviation 0

The non-parametric Kruskal Wallis statistical analysis revealed marginally significant differences among the three gender pairings in the total number of LREs (X^2=5.722, p=0.057) and the total number of meaning related LREs (X^2=5.773, p=0.056) and a statistically significant difference in the number of target-like resolved meaning related LREs (X^2=6.361, p=0.042). However, these differences emerged only when comparing mixed-gender dyads with the MM and the FF dyads but not between same-gender dyads. Note that although it was the case that all form-related LREs were produced by MM dyads, the statistical analysis did not reveal a difference when compared to the production of form-related LREs by the FF dyads. Mixed-gender dyads produced fewer LREs than MM (Z=-2.066, p=0.039) and FF dyads (Z=-2.089, p=0.037); fewer meaning-related LREs than MM (Z=-1.957, p=0.050) and FF (Z=-2.089, p=0.037) dyads; fewer resolved meaning-related LREs than MM (Z=-2.058, p=0.040) and FF (Z=-1.970, p=0.049) dyads; and fewer target-like resolved meaning-related LREs than FF dyads (Z=-2.189, p=0.029).

To answer the second research question, within-group analyses were carried out for each dyad type (MM, FF and mixed-gender pairings). Within group comparisons indicated there were differences within MM (X^2=84.545, p=0.000), FF (X^2=41.020, p=0.000) and mixed-gender dyads (X^2=29.172, p=0.001). All three types of dyads produced more meaning-related LREs than form-related LREs (MM (Z=-3.320, p=0.001), FF (Z=-2.023, p=0.043) and mixed-gender (Z=-2.032, p=0.042)). However, while further differences were observed between meaning and form LREs in both same-gender dyads, no such differences emerged in the mixed-gender dyads. More specifically, MM dyads produced more resolved than unresolved meaning-related LREs (Z=-2.528, p=0.011) and marginally they also produced more resolved than unresolved form-related LREs (Z=-1.841, p=0.066). Nevertheless, the number of meaning- and form-related LREs resolved in a target-like manner did not significantly differ from those resolved in a non-target-

like manner. When comparing the number of target-like resolved form with target-like meaning LREs, a marginally significant difference emerged in favour of meaning-related LREs (Z=-1.910, p=0.056). The comparison between meaning- and form-related LREs as regards non-target-like episodes also yielded a statistically significant difference in favour of meaning-related LREs (z=-2.816, p=0.005). Within the FF dyads, the analysis showed a significantly higher mean of resolved over unresolved meaning-related LREs (Z=-2.032, p=0.042). Similarly, in FF dyads the contrast between target-like resolved and non-target-like resolved LREs did not reach statistical significance.

5. Discussion

The first research question focused on the existence of any differences among MM, FF and mixed-gender dyads in the number, types and outcomes of LREs produced by 10-12-year-old children. The results revealed that same-gender pairings produced a significantly higher number of LREs and resolved LREs than mixed-gender dyads. This seems to show that children as young as 10-12 benefit from working in same-gender dyads more than in mixed-gender dyads. This result has also been reported for L1 (Fabes, Martin, & Hanish, 2003) and L2 children (Azkarai & Imaz Agirre, 2017) but not in those studies that have investigated LRE production with adult L2 learners. In this sense, Azkarai (2015b), Azkarai and García Mayo (2012) and Ross-Feldman (2005, 2007) reported no differences in the occurrence and nature of LREs between mixed- and same-gender dyads during interaction among adults. However, when it comes to resolution of LREs, Azkarai (2015b: 25) reported that mixed-gender dyads had more opportunities to resolve LREs than same-gender dyads, perhaps "because males seemed to be more decisive than females during interactive work". Additionally, Ross-Feldman (2005, 2007) also reported a difference in resolution in that mixed-gender dyads were more favourable for males while same-gender dyads seemed to be more beneficial for females.

As regards the outcome of the resolution of LREs (i.e. whether the LREs were resolved in a target-like or in a non-target-like way), the rate of target-like resolutions was always below 50% in the three pairings (MM: 30.51%, FF: 46.88% and Mixed: 27.27%) despite having been asked to focus on accuracy in the instructions provided. In other words, the inherent focus of the task overruled the framing of the task (Philp, Walter, & Basturkmen, 2010). As reported in previous studies with adults (Adams, 2006; Adams & Ross-Feldman, 2008; García Mayo & Azkarai, 2016; Niu, 2009; Payant & Kim, 2019), oral tasks are more immediate and offer more communicative pressure to the learner, which may limit opportunities for greater accuracy. However, unlike adults and adolescents from previous investigations (Lasito & Storch, 2013; Niu, 2009), children from the present investigation showed even more difficulties in rapidly retrieving and verbalizing their explicit knowledge which could lead them to solve

their gaps in a more target-like way. This might indicate that negotiating successfully is a skill that needs a longer time to develop in young learners (see Gallardo-del-Puerto & Martínez-Adrián, in press in this respect).

The examination of the results also revealed a higher number of target-like resolved LREs in FF dyads than in mixed-gender dyads but no differences between mixed-gender and MM dyads. Even if the rate of target-like resolutions for meaning-related LREs is always below 50% in the three pairings (MM: 30.51%, FF: 46.88% and Mixed: 27.27%), a slight advantage is observed on the part of FF over mixed-gender pairings. In fact, the examination of the descriptive means shows that target-like resolutions are more common than non-target-like episodes in the FF dyads, while the opposite trend is obtained in the MM and mixed-gender pairings, a result which will also be discussed in connection to the second research question.

Moreover, the between-group analysis indicated the inexistence of statistically significant differences between MM and FF dyads in incidence, nature and resolution of LREs. Thus, there is little impact of the variable 'gender' as also reported in Azkarai and Imaz Agirre (2017) with EFL children and Oliver (2002) with ESL children and Azkarai (2015b) with EFL adults. However, the results partially support Ross-Feldman (2005, 2007) who found no differences in incidence but in resolution as more resolved LREs were attested in the adult male learners examined in her sample.

With respect to the second research question which inquired into the type of LREs (nature and outcome) which are more common in each dyad group, the analysis indicated that in the three dyad groups, and as regards the nature of LREs, learners significantly produced more meaning-related than form-related LREs. The task administered, an oral task, is a communication-oriented task with an inherent immediate nature, and to satisfy its demands, these young learners are in the need of key vocabulary to move it forward (García Mayo & Azkarai, 2016; García Mayo & Imaz Agirre, 2019; Payant & Kim, 2019; Swain & Watanabe, 2013). This contrasts with tasks that offer the learners more time to think and to reflect on their production, where young learners have been found to focus both on meaning and form (Gallardo-del-Puerto & Martínez-Adrián, in press; García Mayo & Imaz Agirre, 2019). Notwithstanding, it is worth mentioning that only male-male pairings were found to produce form-related LREs in addition to meaning-focused LREs in this oral task. We could tentatively argue that males discuss a wider range of issues not just limited to lexis, whereas females are more concerned with conveying the message in a task whose primary focus is meaning. In this vein, this finding aligns with prior L1 research, where males have been reported to discuss many topics briefly in

matched-gender dyads (Tannen, 1990)[6]. In addition, other investigations on the use of communication strategies with same-age learners have also revealed how girls report being less risky than boys when communicating, which shows a strong preference for efficiency so as to get their message across without communication breakdowns (see Basterrechea, Martínez-Adrián, & Gallardo-del-Puerto, 2017).

In terms of resolution, the finding that young male and female learners produce more resolved than unresolved LREs in matched-gender dyads may indicate that children generate more opportunities to learn when working in same-gender dyads. This result aligns with L1 acquisition studies according to which same-gender peer preferences in children become stronger over time (Hanish & Fabes, 2014). Hanish and Fabes (2014) suggest that the preference for same-gender dyads leads to fewer opportunities for male-female interaction and hence, fewer opportunities for males and females to learn from each other. This observation is reflected in our study in that mixed-gender dyads are not as effective as same-gender dyads when it comes to resolving LREs.

But even if resolved LREs were significantly higher than unresolved LREs in matched-gender pairings, there was a low rate of target-like resolved LREs (30.51% in MM and 46.88% in FF). In addition to the low rate of target-like resolution in MM and FF dyads, it is also important to highlight the fact that although the difference between both dyads did not reach statistical significance, we can observe a higher mean number of target-like episodes over non-target-like ones in FF dyads, whereas in MM non-target-like episodes outnumber target-like resolutions. This finding might tentatively indicate that FF dyads could be more oriented to upsurge the correctness of their language discussion outcomes in a task that demands lower levels of accuracy. In this regard, it seems as if these FF dyads conform to the instructions provided to focus on accuracy to a higher extent than boys, a finding which seems to be in line with research in education (see girls' advantage in terms of the social and behavioural skills that are valuable in producing higher levels of academic performance (DiPrete & Buchman, 2013)).

6 This idea seems to be reinforced in that these very same learners (either in male-male or female-female dyads), when tested in another study in which they were asked to perform a speaking+writing task that seems to me more powerful in increasing learners' attention to formal issues (see Gallardo-del-Puerto & Martínez-Adrián, in press), were found to produce form-related episodes. So, in the light of these findings, we cannot argue that males are more oriented to form (i.e. grammar, pronunciation issues). Nevertheless, we will have to await other studies that analyse a wider range of tasks to confirm this trend.

6. Conclusion

This study has revealed that type of pairing affects LRE production and resolution as more LREs were initiated and resolved in matched-gender dyads than in mixed-gender dyads. However, gender-pairing did not have an impact on the outcome of the resolution, as a still low rate of targetlikeness was obtained in mixed- and matched-gender young dyads. In addition, gender was not a discriminating factor when comparing same-gender dyads, even though FF dyads seem to be more concerned with getting the message across and more oriented to higher accuracy in the resolutions of meaning-related episodes as the descriptive means evinced, findings which need to be confirmed in further studies with larger samples of young learners and a more balanced distribution of dyad types.

In the light of the results, two main pedagogical implications may be drawn for the EFL class. Younger learners may benefit more from working in matched-gender dyads than in mixed-gender dyads, at least in this type of communicative tasks that enhance collaborative dialogue. Second, the inclusion of these oral tasks in the curriculum promotes language focus for young learners and in particular the collaborative construction of meaning. But despite the potential of the adoption of these oral tasks in the curriculum for negotiation of meaning, learning opportunities for young learners should be maximized in this task by the implementation of measures aimed at increasing the accuracy resolution rate. In this respect, ventures already tested with the same learners such as the possibility of edition of their oral production in a similar task as the one employed in the present study have been found to promote more target-like resolutions of meaning-related episodes (Gutiérrez-Mangado & Basterrechea, forthcoming; Martínez-Adrián & Gallardo-del-Puerto, forthcoming). Likewise, other task options and modalities should be considered to boost accuracy (see Martínez-Adrián, Gutiérrez-Mangado, Gallardo-del-Puerto, & Basterrechea, in press). In this regard, the incorporation of focused tasks accompanied by appropriate training conditions could enhance metalinguistic awareness, facilitate a greater noticing of gaps and in turn, a better development of language accuracy (Bouffard & Sarkar, 2008; Shak & Gardner, 2008).

For future research, taking into account that these results cannot be fully generalizable to other contexts, it would be desirable to conduct this type of research in other cultural settings, as cultural norms have been found to interact with other social factors such as gender. In addition, a wider range of tasks and modalities, as well as the degree of elaboration of LREs (Niu, 2009) should be explored. By investigating more and less focused tasks in which writing components are added that usually lead to greater noticing of grammar features as well as oral tasks in which learners are offered the possibility of editing their production, we will be able to look into whether

similar trends reported in this study as regards gender-pairings are obtained. Thus, research on the interplay between gender-pairing and task-modality is clearly advocated so as "to inform policy makers and maximise children's opportunities for learning" (García Mayo, 2017: xv). Likewise, the administration of tailor-made posttests could contribute to the study of collaborative dialogue on language development. Similarly, (pseudo)longitudinal studies could also shed more light on the provision of learning opportunities for young learners throughout time. In this respect, we will be able to delve into whether the better performance observed in matched-gender dyads persists during these school years.

7. Acknowledgements

The authors would like to acknowledge the grants awarded by the Spanish Ministry of Economy and Competitiveness (FFI2016-74950-P) (AEI/FEDER/UE) and the Basque Government (IT904-16). The authors are also grateful to the school, the teachers and students for their collaboration in the research project.

8. References

Adams, R. (2006). L2 tasks and orientation to form: A role for modality? *ITL International Journal of Applied Linguistics, 152*, 7-34.

Adams, R., & Ross-Feldman, L. (2008). Does writing influence learner attention to form? In D. Belcher & A. Hirvela (Eds.), *The Oral-literate Connection. Perspectives on L2 Speaking, Writing, and other Media Interactions.* (pp. 243-265). Ann Arbor, MI: The University of Michigan Press.

Alcón, E. (1994). The role of participation and gender in non-native speakers' classroom interaction. *Working Papers on Language, Gender, and Sexism, 4*(1), 51-68.

Alcón, E., & Codina, V. (1996). The impact of gender on negotiation and vocabulary learning in a situation of interaction. *International Journal of Psycholinguistics, 12*, 21-35.

Aries, E.J. (1976). Interaction patterns and themes of male, female, and mixed groups. *Small Group Behaviour, 7*(1), 7-18.

Azkarai, A. (2013). *Gender and task-modality in EFL task-based interaction*. Unpublished PhD thesis., The University of the Basque Country.

Azkarai, A. (2015a). L1 use in EFL task-based interaction: a matter of gender? *European Journal of Applied Linguistics, 3*(2), 159-179.

Azkarai, A. (2015b). Males and females in EFL task-based interaction: does gender have an impact on LREs? *Vigo International Journal of Applied Linguistics, 12*, 9-35.

Azkarai, A., & Imaz Agirre, A. (2017). Gender and age in child interaction in an EFL CLIL context: An exploratory study. In M. P. García Mayo (Ed.), *Learning Foreign Languages in Primary School: Research Insights*. (pp. 103-123). Clevedon, UK: Multilingual Matters.

Azkarai, A., & García Mayo, M. P. (2012). Does gender influence task performance in EFL? Interactive tasks and language related episodes. In E. Alcón & M. P. Safont Jordá (Eds.), *Language Learners' Discourse across L2 Instructional Settings*. (pp. 249-278) Amsterdam: Rodopi.

Azkarai, A., & García Mayo, M. P. (2015). Task modality and L1 use in EFL oral interaction. *Language Teaching Research, 19*, 550-571.

Basterrechea, M., Martínez-Adrián, M., & Gallardo-del-Puerto, F. (2017). Gender effects on strategic competence: A survey study on compensatory strategies in a CLIL context. *Estudios de Lingüística Aplicada (ELIA)*, *17*, 47-70.

Birjandi, P., & Tabatabaei, O. (2010). Examining the impact of gender on task-based conversational interactions among EFL learners: The case of recasts. *Curriculum Planning, Knowledge and Research in Educational Sciences,* 7(26), 35-64.

Bohn, E., & Stutman, R. (1983). Sex-role differences in the relational control dimension of dyadic interaction. *Women's Studies in Communication, 6*, 96-104.

Bouffard, L. A., & Sarkar, M. (2008). Training 8-year-old French immersion students in metalinguistic analysis: An innovation in form-focused pedagogy. *Language Awareness*, *17*(1), 3-24.

Cenoz, J. (2009). *Towards Multilingual Education: Basque Educational Research from an International Perspective*. Bristol: Multilingual Matters.

DiPrete, T. A., & Buchman, C. (2013). *The Rise of Women. The Growing Gender Gapa in Education and What it Means for American Schools*. New York: Russell Stage Foundation.

Fabes, R. A., Martin, C. L., & Hanish, L. D. (2003). Young children's play qualities in same-, other-, and mixed-sex peer groups. *Child Development*,*74*, 921–932.

Gallardo-del-Puerto, F., & Martínez-Adrián, M. (In press). Task modality effects on young learners' language related episodes in collaborative dialogue. *RESLA*.

Gallardo-del-Puerto, F., Basterrechea, M., & Martínez-Adrián, M. (2020). Target language proficiency and reported use of compensatory strategies by young CLIL learners. *International Journal of Applied Linguistics, 30*, 3-18.

García Mayo, M. P. (2017). Introduction. In M. P. García Mayo (Ed.), *Learning Foreign Languages in Primary School.* (pp. xiii-xxiii). Bristol: Multilingual Matters.

García Mayo, M. P., & Azkarai, A. (2016). EFL task-based interaction: Does task modality impact on language-related episodes? In M. Sato & S. Ballinger (eds) *Peer Interaction and Second Language Learning: Research Agenda and Pedagogical Potential.* (pp.241–266). Amsterdam: John Benjamins.

García Mayo, M. P., & Hidalgo Gordo, M. A. (2017). L1 Use among young EFL mainstream and CLIL learners in task-supported interaction. *System, 67*, 132–145.

García Mayo, M.P., & Imaz Agirre, A. (2019). Task modality and pair formation method: Their impact on patterns of interaction and LREs among EFL primary school children. *System, 80*, 165-175.

García Mayo, M. P., & Lázaro Ibarrola, A. (2015). Do children negotiate for meaning in task-based interaction? Evidence from CLIL and EFL settings. *System,* 54, 40-54.

Gass, S. & Varonis, E. (1986). Sex differences in non-native speaker/nonnative speaker interactions. In R.R. Day (Eds.), *"Talking to Learn": Conversation in Second Language Acquisition.* (pp.327-351). Cambridge, MA: Newbury House.

Gu, L. (2013). Language and gender: differences and similarities. *Proceedings of the International Conference on Advances in Social Science, Humanities, and Management (ASSHM-13)*, 248-251.

Gutiérrez-Mangado, M. J., & Basterrechea, M. (forthcoming). Investigating the effect of task modality on the production of LREs by young learners.

Hanish, L. D., & Fabes, R. A. (2014). Peer socialization of gender in young boys and girls. *Encyclopedia on Early Childhood Development. Montreal, Quebec: Centre of Excellence for Early Childhood Development and Strategic Knowledge Cluster on Early Child Development*, 1-4. http://www.child-encyclopedia.com/sites/default/files/textes-experts/en/2492/peer-socialization-of-gender-in-young-boys-and-girls.pdf

Holmes, J. (1994). Improving the lot of female language learners. In J. Sunderland *(ed) Exploring Gender: Questions and implication of English Language Education.*, (pp. 156-162). London: Prentice Hall.

House, S., & Scott, K. (2009). *Sparks 1*. London: Richmond.

Lasito, & Storch, N. (2013). Comparing pair and small group interactions on oral tasks. *RELC Journal, 44*, 361-375.

Lázaro Ibarrola, A., & Azpilicueta Martínez, R. (2015). Investigating negotiation of meaning in EFL children with very low levels of proficiency. *International Journal of English Studies, 15*, 1-21.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (eds) *Handbook of Second Language Acquisition.* (pp. 413-468). San Diego, CA: Academic Press.

Maltz, D. N., & Borker, R. A. (1982). A cultural approach to male-female miscommunication. In J. Gumperz (Eds.), *Language and Identity.* (pp.195-216). Cambridge: Cambridge University Press.

Martínez-Adrián, M. (2020). ¿Los juntamos? A study of L1 use in interactional strategies in CLIL vs. non-CLIL primary school learners. *International Review of Applied Linguistics in Language Teaching (IRAL), 58*(1), 1-27.

Martínez-Adrián, M., & Gallardo-del-Puerto, F. (forthcoming). Task modality: The control of accuracy and edition in output.

Martínez-Adrián, M., Gutiérrez-Mangado, M. J., Gallardo-del-Puerto, F., & Basterrechea, M. (In press). Language-related episodes by young CLIL learners: A review of task modality effects. *Language Teaching for Young Learners*.

McWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Mehta, C. M., & Strough, J. (2009). Sex segregation in friendships and normative contexts across the life span. *Developmental Review, 29*, 201-220.

Mehta, C. M., & Strough, J. (2010). Gender segregation and gender-typing in adolescence. *Sex Roles, 63*, 251-263.

Niu, R. (2009). Effect of task-inherent production modes on EFL learners' focus on form. *Language Awareness,* 18 (3-4), 384-402.

Oliver, R. (2002). The patterns of negotiation for meaning in child interactions. *The Modern Language Journal, 86*(1), 97-111.

Oliver, R., & Azkarai, A. (2017). Review of child second language acquisition (SLA): Examining theories and research. *Annual Review of Applied Linguistics, 37*, 62-76.

Pavlenko, A. (2001). "How am I to become a woman in an American vein?": Negotiation of gender in second language learning. In A. Pavlenko, A. Blackledge, I. Piller, & M. Teutsch-Dwyer (Eds.), *Second Language Learning, Multilingualism, and Gende*r. (pp. 133-174). Berlin: Mouton De Gruyter.

Pavlenko, A., & Piller, I. (2008). Language education and gender. In S. May & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education*. (pp. 57-69). New York: Springer Science+Business Media LLC.

Payant, C., & Kim, Y. J. (2019). Impact of task modality on collaborative dialogue among plurilingual learners: a classroom-based study. *International Journal of Bilingual Education and Bilingualism, 22*(5), 614-627.

Philp, J., Walter, S., & Basturkmen, H. (2010). Peer interaction in the foreign language classroom: What factors foster a focus on form? *Language Awareness, 19*(4), 261-279.

Pica, T., Holliday, L., Lewis, N. E., Berducci, D., & Newman, J. (1991). Language learning through interaction: What role does gender play? *Studies in Second Language Acquisition, 13*, 343 - 376.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction and research. In G. Crookes & S. M. Gass (Eds.), *Tasks and Language Learning: Integrating Theory and Practice.* (pp. 9-34). Clevedon: Multilingual Matters.

Pica, T., Kang, H., & Sauro, S. (2006). Information gap tasks: Their multiple roles and contributions to interaction research methodology. *Studies in Second Language Acquisition, 28*, 301-338.

Pinter, A. (2011). *Children Learning Second Languages*. Basingstoke, U.K: Palgrave Macmillan.

Pladevall Ballester, E., & Vraciu, A. (2020). EFL child peer interaction: Measuring the effect of time, proficiency pairing and language of interaction. *Studies in Second Language Learning and Teaching, 10*(3), 449-472.

Ross-Feldman, L. (2005). *Task-based interactions between second language learners: Exploring the role of gender*. Unpublished PhD thesis, Georgetown University.

Ross-Feldman, L. (2007). Interaction in the L2 classroom: Does gender influence learning opportunities? In A. Mackey (Eds.), *Conversational Interaction in Second Language Acquisition: A Collection of Empirical Studies.* (pp.52-77). Oxford: Oxford University Press.

Sadker, M., & Sadker, D. (1994). *Failing at Fairness: How our Schools Cheat Girls*. New York: Simon & Schuster.

Shak, J., & Gardner, S. (2008). Young learner perspectives on four focus-on-form tasks. *Language Teaching Research*, *12*(3), 387–408.

Shehadeh, A. (1994). Gender differences and second language acquisition. *Research Journal of Aleppo University (Arts and Humanities Series), 26*, 73 - 98.

Shehadeh, A. (1999). Gender differences and equal opportunities in the ESL classroom. *ELT Journal 53*, 256-261.

Spolsky, B. (1989). *Conditions for Second Language Learning: Introduction to a General Theory*. Oxford: Oxford University Press.

Sunderland, J. (2000). Issues of language and gender in second and foreign language education. *Language Teaching,* 33(4), 203-223.

Swain, M. (1998). Focus on form through conscious reflection. In by C. Doughty & J. Williams (ed) *Focus on Form in Classroom Second Language Acquisition.* (pp. 64-81). Cambridge: Cambridge University Press.

Swain, M., & Watanabe, Y. (2013). Languaging: Collaborative dialogue as a source of second language learning. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics.* Oxford: Wiley-Blackwell.

Tannen, D. (1990). *You Just Don't Understand. Women and Men in Conversation*. New York: William Morrow.

Tannen, D. (1994). *Gender and Discourse*. New York: Oxford University Press.

Van Lier, L. (1988). *The Classroom and the Language Learner: Ethnography and Second-Language Classroom Research*. London: Longman.

West, C., & García, A. (1988). Conversational shift work: A study of topical transitions between women and men. *Social Problems, 35*(5), 551-575.

West, C., & Zimmerman, D. (1983). Small insults: A study of interruptions in cross-sex conversations between unacquainted persons. In B. Thorne, C. Kramarae & N. Henley (Eds.), *Language, Gender and Society* (pp.102-117). Rowley, Mass: Newbury House.

Zalbide, M., & Cenoz, J. (2008). Bilingual education in the Basque Autonomous Community: Achievements and challenges. In J. Cenoz (Ed.), *Teaching through Basque.* (pp.5-20). Clevedon: Multilingual Matters.

Zimmerman, D., & West, C. (1975). Sex roles, interruptions and silences in conversation. In B. Thorne & N. Henley (Eds.), *Language and Sex: Difference and Dominance.* (pp.105-129). Rowley, Mass: Newbury House.

Appendix 1

Source: https://www.guiassantillana.com/ingl%C3%A9s/sparks/

Second language acquisition of Spanish prosody by Chinese speakers: Nuclear contours and pitch characteristics

Peizhu Shang
Department of Catalan Philology and General Linguistics
University of Barcelona, Spain
pshangsh7@alumnes.ub.edu

Wendy Elvira-García
Department of Spanish Language and General Linguistics
UNED, Spain
welvira@flog.uned.es

Abstract

Despite the increasing number of studies in L2 prosody, little research has been carried out on the Chinese-Spanish language pair. This article sets out to examine the L2 acquisition of nuclear contours and pitch implementation details of Spanish spoken by Chinese speakers. To this end, 555 utterances (produced by 37 informants) were analyzed within an autosegmental-metrical framework, and pitch values were evaluated using long-term distributional (LTD) and pitch dynamism quotient (PDQ) measures. The results suggest a hierarchy of difficulties in acquiring the prosodic features of different sentence types. The most salient intonational error made by the Chinese learners was the tendency to replace low nuclear accents with high/rising tones. Furthermore, the higher pitch level, narrower span, and lower F0 variance found for Chinese speakers lend support to previous hypotheses which proposed a general pitch compression pattern for L2 speech. Nevertheless, with increasing proficiency in Spanish, learners appear to develop more target-like intonation contours and pitch profiles. Finally, gender and stress effects as well as other interactions prove that L2 prosody learning is more complex than previously stated, and is influenced not only by the L1 system and oral competence but is also correlated with some psychological and sociocultural factors.

Key words: Nuclear contour; pitch characteristics; L2 speech acquisition; prosodic transfer; Peninsular Spanish by L1 Chinese speakers.

Resumen

A pesar del creciente número de estudios sobre L2, el par de lenguas chino-español ha sido abordado en escasos estudios. Este artículo examina la adquisición de los contornos entonativos y los detalles de implementación tonal del español por parte de los sinohablantes. Para hacerlo, se han analizado 555 oraciones producidas por 37 informantes usando la teoría métrico-autosegmental, y se han evaluado acústicamente sus valores tonales a través de las medidas de distribución a largo plazo (LTD) y coeciente de dinanismo de tono (PDQ). Los resultados sugieren una jerarquía de dificultad en la adquisición de los patrones entonativos relacionada con los diferentes tipos oracionales. Además, el error entonativo más destacado para los sinohablantes es la tendencia a reemplazar los acentos nucleares bajos por tonos altos/ascendentes. Además, nuestros resultados apoyan las hipótesis que proponen una tendencia general de compresión tonal en la L2, ya que se ha documentado un nivel tonal más alto, un rango más reducido y una variación menor de F0 en los sinohablantes que en el grupo control. Sin embargo, los aprendices parecen desarrollar contornos entonativos y perfiles tonales más parecidos al objetivo a medida que aumenta la competencia en L2. Finalmente, el efecto del género y acento, así como otras interacciones, demuestran que la adquisición de la prosodia es más compleja de lo expuesto anteriormente, y que se ve influenciada no solo por el sistema de la L1 y las competencias orales, sino también por factores psicológicos y socioculturales.

Palabras clave: Contorno nuclear; característica tonal; adquisición del habla de L2; tranferencia prosódica; español peninsular por los sinohablantes.

1. Introduction

Prosody in second language acquisition (SLA) research has long been promoted through a monolingual linguistic ideology. The vast majority of SLA studies on phonetics and phonology, to date, overtly or covertly tend to treat native speakers as a yardstick for second language (L2) learning. In this regard, L2 learners are defined as failed and deficient monolinguals that will not achieve a native proficiency because of their unchangeable conditions at birth (Cook, 2012; Ortega, 2010, 2011, 2014). However, since the mid-1990s, there has been continual criticism against this dominant paradigm and the concept of monolingualism as a default norm for SLA. For instance, Ortega pointed that the major deleterious consequence of the monolingual view is that it unethically turns bilinguality into an invisible reality by "erasing bilinguals' other language competence from analysis" (2010: 56). To overcome this pervasive monolingual bias, a new trend of bi/multilingual turn has been developed in recent years (May, 2013; Ortega, 2010, 2011, 2014; Valdés, Poza & Brooks, 2015). The creativity

of this new paradigm is that it conceives SLA as a dynamic and constantly interactive state between the first (L1) and the second language, rather than the sum of two clearly separate monolinguals (May, 2013). Besides, instead of examining SLA development based on native speakers, it proposes to investigate learners' total language repertories, and utilize fully-developed bilinguals or successful L2 users as a new empirical baseline for L2 learning (Cook, 2016; Ortega, 2010, 2014).

Despite the advantage of a bilingual approach portraying a more complete picture for SLA research, it has not been sufficiently worked out to replace the predominant SLA theories, at least in the domain of phonetics and phonology. The main causes of this, according to May (2013), have been correlated with the ideological roots of monolingualism as the normative condition of the discipline. Another possible reason lies in the ambiguity and complexity of correctly defining the mature and experienced L2 users as the reference for SLA (Cook, 2016). Moreover, the lack of integration between theories and practices in different SLA contexts and fields also have stunted the epistemological change toward bilingual turn in SLA (Pandey, 2016). Finally, as mentioned by Pandey (2016), cross-continental examples or typologically distinct language pairs are still needed to explore the global appeal of current frameworks on the bilingual turn. Given these remaining uncertainties in the bilingual turn and the typological distance between Chinese and Spanish prosody (see the discussion in section 2.1), we decided in this study to investigate SLA from the classical view of native speakers. Unlike Spanish and most Romance languages that are intonation languages, Chinese is widely known as a tonal language, with each tone having a different internal pattern of rising and falling pitch contours (Yip, 2002). Arguably, problems would arise when comparing the intonation of Chinese with Spanish, given that the pitch contours of Chinese are highly dependent on the tone type of the sentence-final syllable. Therefore, rather than direct comparisons across learners' language pairs, it seems to be more feasible in our research to interpret the L1 influence on L2 acquisition based on previous empirical studies on Chinese prosody. The L2 transfer to L1 prosody from the opposite direction will be considered for further research.

Cross-language transfer effects have been discussed over the last decades in SLA research and most studies have focused on segmental differences across languages. However, several intonational aspects, for instance, F0 register, pitch range, and intonational pitch patterns are, in essence, more vulnerable to cross-language influences, and therefore more difficult to learn, to interpret, and to investigate in SLA (Mackay, 2000; Atoye, 2005; Mennen, 2014, 2015). These difficulties in L2 prosody are not merely caused by the typological distance between the phonological and phonetic systems of the L1 and the L2, but are also closely linked to the "complexity and multidimensionality" of intonation (Mennen & Leeuw, 2014: 187). Intonation

can interact with other prosodic structural elements (e.g. duration, rhythm, intensity and lexical stress) and signal multiple pragmatic functions in the speech. Therefore, it is difficult to determine whether some intonational variations in the phonetic implementation are either categorical or gradient to the phonological representations (Mennen, 2015; Nolan, 2006). Besides, the divergence of different types of frameworks on prosodic typology (e.g. see the discussion between Beckman & Venditti, 2011; Hyman, 2006, 2012; Jun, 2006; Ladd, 2001) adds to the existing difficulty of describing and characterizing intonational features. This methodological issue remained unsolved in the prosodic field until the advent of the AM (autosegmental-metrical) theory that has allowed researchers to analyze and uniformly compare the intonation systems of many languages.

The broadly recognized autosegmental-metrical framework was developed around the core idea of isolating the categorial phonological elements from its surface phonetic realization (Gussenhoven, 2004; Ladd, 1996; Pierrehumbert, 1980). This separation between the phonological and phonetic dimensions is of great importance, not only for the prosodic investigation of a wide variety of languages (e.g. Hualde, 2003; Face & Prieto, 2007; Estebas-Vilaplana & Prieto, 2008, 2010 for Spanish; Tseng, Huang, & Beckman, 2011 for Mandarin Chinese), but also for cross-language or cross-dialectal comparisons of intonation (e.g. English and Spanish comparison by Bowen, 1956 and Vilaplana, 2008; Mandarin and English by Crosswhite & McDonough, 2000; Spanish and Catalan by Vilaplana, 2008; Majorcan and Minorcan Catalan by Payà & Vanrell, 2005). According to the AM theory, intonational pitch contours are phonologically comprised of two types of tonal units: pitch accents and boundary tones, which are realized as either a high or low tone attached to metrically prominent syllables or the edges of prosodic phrases. Different languages may differ typologically in the inventory of pitch accents, boundary tones and combinations, using strategies of the structural elements according to the text structure (Graham & Post, 2018; Ladd, 1996; Mennen, 2015). Beyond this, the surface phonetic shapes (e.g. pitch height, pitch span, alignment, speech rhythm, etc.) of the intonation primitives (basic components) may also vary to different degrees because of different conventions of pitch implementation across language communities.

Furthermore, in the recently developed L2 Intonation Learning Theory (LILt), Mennen (2015) suggested that there can be some cross-language intonation differences in the semantic and frequency dimensions, not merely in the systematic (or 'phonological') and realizational (or 'phonetic') aspects. Concerning deviations in the semantic dimension, Chinese learners of L2 English were found to have difficulties in signaling new information (Juffs, 1990), marking contrastive stress (Wennerstrom, 1998) and prominence relationships (McGory, 1997) in a native-like way, even

having reached a high level of proficiency in the L2. Besides, significant contrasts were observed in the frequency of usage of phonological tones in the L1 and the L2, probably due to the transfer of the first language (see examples in Backman, 1979; Hewings, 1995; Jilka, 2000; Mennen et al., 2010). Based on this multi-dimension modeling, L2 intonation research is expected not only to examine learners' phonetic and phonological deviations from the target language but also to shed new light on their ability to use appropriately different categorical types and their ability to use intonation to signal different functional meanings in the cross-language speech.

Overall, the current study was undertaken based on the deep phonological understanding of intonation (AM) and previous evidence developed in the L2 prosody learning model (LILt). It was of interest to investigate the systematic acquisition of the intonation contours and pitch implementation details in five question types with different linguistic meanings in Spanish by L1 speakers of Mandarin Chinese and Peninsular Spanish. Furthermore, this study examined the correlation between different pitch variables and compares prosodic performance across different proficiency levels, question types, stress positions, and gender. The rest of the paper is structured as follows: Section 2 describes the intonation system of Peninsular Spanish and Mandarin Chinese and reviews the existing literature on pitch range implementation. Beyond this, the cross-language differences between L1 and L2 prosody will be introduced in this section. Next, Sections 3 and 4 describe the present study with the specific research questions and the methodology. Sections 5 and 6 report the experimental results and discuss the main findings of the work. Finally, in Section 7, we present the conclusions, the pedagogical implications, and some potential limitations of the current study.

2. Literature review

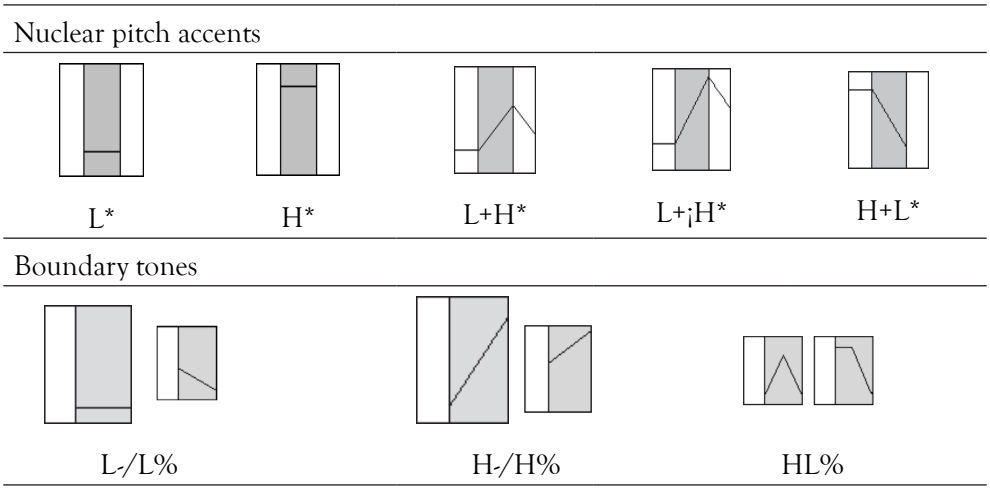
2.1. Intonation contours in L1 and L2

As Spanish is an intonation language, there is a general consensus on the use of ToBI (Tones and Break Indices) labeling systems for the description of intonational phonology. ToBI conventions distinguish two types of tonal events (pitch accents and boundary tones) and describe the intonational patterns by means of two tones that indicate prosodic level (e.g. low is L, high is H). If the tone is associated with a prominent or metrically strong syllable, the letter is followed by a star (*), for example, L* stands for a low stressed syllable. In Spanish, if the prominent syllable has more than one tone attached, the tones are linked with a plus sign and the star indicates the position of the stress. For example, L+H* signals a rising stressed syllable, whereas

L*+H signals a low stressed syllable and a high poststressed syllable. The group of tones associated to a strong syllable is called a pitch accent. If the tone is associated with the edge of an intonation group, it is marked with a percentage sign (%) when the intonation group is final (signaling an intonation phrase), or with a dash (-) when it is non-final (signaling an intermediate phrase).

Based on the ToBI framework, 2 monotonal pitch accents (L*, H*), 3 bitonal pitch accents (L+H*, L+¡H*, H+L*), 2 monotonal boundary tones (L-/L%, H-/H%) and 2 bitonal boundary tones (LH%, HL%) (see the schematic representations in Figure 1) have been proposed for the nuclear configuration of questions in Peninsular Spanish (Estebas-Vilaplana & Prieto, 2010; Estebas-Villaplana, 2008; Face & Prieto, 2007; Hualde & Prieto, 2015; Prieto & Roseano, 2019). While the monotonal accent L* is commonly found in the nuclear position of interrogative modality (e.g. in the nuclear location of information-seeking *yes-no* questions, information-seeking *wh*-questions, confirmation-seeking tag questions as well as the final constituent of disjunctive questions), the high pitch accent H* often appears in the nuclear position of echo and rhetorical questions (Estebas-Vilaplana & Prieto, 2010). With regard to the bitonal accents L+H* and L+¡H*, these are frequently linked to obviousness meanings or some nuance of surprise and exclamation, and normally appear in statements of the obvious, exclamative statements and various counterexpectational questions of Peninsular Spanish. The phonetic distinction between the two accent types typically lies in the F0 scaling, with L+¡H* exhibiting a step rise into the highest peak of the utterance. In contrast with the two previous rising accents, in Peninsular Spanish a falling pattern (e.g. the H+L* L% found for confirmation-seeking *yes-no* question) is frequently used to express "the degree of commitment that a speaker has in the truth-value of a proposition" (which technically has been called the epistemicity) (Roseano, González, Borràs-Comes, & Prieto, 2016: 3). Nevertheless, the epistemicity encoded by the intonation strategy is generally considered to be weaker than that marked with confirmative tags, for example, *¿no?* (no?), *¿verdad?* (is it?), *¿no es cierto?* (isn´t it?). In Spanish, these syntactic structures are typically pronounced with a low nuclear accent L* followed by a high-rising boundary tone H%. Aside from tag questions, the H% or H- edge tone is also attested in information-seeking *yes-no* questions and the non-final constituents of disjunctive questions. Conversely, the low-falling boundary tone L% is usually observed at the end of information-seeking *wh*-questions and disjunctive questions as well as confirmation-seeking *yes-no* questions.

Figure 1. Schematic representations of the pitch accents and boundary tones of Peninsular Spanish questions presented in our study (adapted from Estebas-Vilaplana & Prieto, 2010: 19-21)



In Mandarin Chinese, however, the realization of intonation types differs strongly from intonation languages such as English and Spanish, where pitch contours are used to convey linguistic meanings and communicative purposes (Prieto & Roseano, 2019) although it is worth noting that pitch variation is typically accompanied by changes in other prosodic features, like duration, amplitude, or voice quality. As a tone language, on the one hand, the key issue in Chinese prosody research is concerned with the interaction between tone and intonation. It has long been established that Mandarin Chinese has four lexical tones and one neutral tone in its inventory, namely, tone1 (high), tone2 (rising), tone3 (falling-rising), tone4 (falling), and tone0 (phonological neutral). Though both tone and intonation use F0 as a primary cue, the functional load of F0 in Chinese lies mostly in word distinction (Chen & Gussenhoven, 2008; Yuan & Shih, 2004). Thus, it is expected that intonational meanings in Mandarin Chinese might be recognized through the modification of other prosodic indicators (normally, pitch range, pitch level, and duration), due to the restriction in manipulating the F0 curve (Chen & Gussenhoven, 2008). On the other hand, despite the growing body of knowledge on tone-intonation interaction (Chao, 1968; Chen & Gussenhoven, 2008; Girding, Zhang, & Svantesson, 1983; Kochanski & Shih, 2003; Wu, 1982; Xu, 2005, 2015; Xu & Wang, 2001; Yuan, 2004, 2006; Yuan & Shih, 2004), there seems to be no consensus as to the formulation of a general framework for Chinese prosody. Various models have been proposed to investigate the Chinese intonation mechanism, the most influential ones being the following: the PENTA model (Xu, 2005; Xu & Wang, 2001), the STEM-ML model (Kochanski & Shih, 2003) and the

StaticTarget model (Shih, 1988). Although models differ in the way they interpret and analyze intonation and its interaction with lexical tones, it is generally accepted that the intonation of Chinese questions and statements diverges increasingly at the sentence-final location. These observations are generally in line with the ToBI-based intonation school who tends to differentiate intonation types through the final nuclear configuration. Nevertheless, the high (H-/H%) and low (L-/L%) boundary labels used in ToBI systems were found to be problematic for the prosodic annotation of Mandarin Chinese (e.g. Pan-ToBI or C-ToBI system), particularly in the intonation of questions, due to the fact that the surface F0 contour of Chinese questions could have either a rising or falling end depending on the tonal identity of the last syllable

The typological differences of prosodic structure and intonation mechanism between the tone languages and the intonation languages could pose significant challenges for L2 learners of tone languages over the course of learning an intonation language. In the last decades, with the growing body of literature on L2 prosody acquisition, various intonation errors have been reported for Chinese learners of L2 English. For instance, it has been found that Chinese speakers of English tend to employ systematically high-level tones (H*) in target nuclear accents where the phonetic realization is characterized by a low plateau (henceforth L*) during the stressed syllable (Barto, 2015; Hong, 2012; Ji, 2010; Ji et al., 2009, 2012; Shao et al., 2011; Xu, 2009). The tendency to use high-level tones in the L2 nuclear position has been mostly correlated with the lack of a steady low tone (L*) in the tonal inventory of Chinese (Ji et al., 2009; Yuan et al., 2019). Additionally, the intonation stress in Chinese is achieved by mainly raising the H target tone rather than changing the low tone. Therefore, as suggested by Ling (2003) and Jia, Xiong & Li (2005), Chinese learners may unconsciously transfer the high tone strategy from their L1 to the L2 target intonation patterns when speaking an intonation language like English. Compared to the numerous findings regarding L2 English intonation produced by Chinese speakers, little research has been carried out in the case of language contact between Spanish and Chinese. To our knowledge, the only empirical investigations carried out have been those by Cortés Moreno (1997, 1999, 2001, 2004, 2005) and Liu (2005) on the L2 acquisition of Spanish prosody by Mandarin-speaking Taiwanese students. Specifically, Cortés Moreno (1997, 2004) found that Taiwanese students had considerably more problems producing the L2 Spanish intonation than perceiving it, even advanced learners who had shown native-like performance in perceiving the target intonation patterns. Moreover, studies by Cortés Moreno (2001, 2005) seem to suggest a hierarchy of difficulties in the acquisition of L2 Spanish intonation types whereby *yes-no* questions were the most difficult pattern to learn in comparison to statements and *wh*-questions. This ranking of learning difficulties holds essential implications for L2 prosody teaching. However, it is interesting to note that the said

proposal was the opposite of the sequence put forth by Yuan et al. (2019). In this latest study, statements were reported to be the most difficult L2 patterns for Chinese students due to their less perceptually salient target pitch movements compared to *yes-no* questions and requests, which involve more H-L contrasts.

2.2. Cross-language research of pitch range variation

Aside from the phonological system, languages may also differ typologically in the way they use pitch to phonetically implement the categorical elements. According to Ladd (1996), Cruttenden (1997) and Gussenhoven (2004), there are two main types of pitch range variation, namely, level and span. Whereas pitch level, or "register" has been referred to as the overall height of a speaker's voice (Cruttenden, 1997), span, or "F0 range" has been described as a speaker's "excursion size" of range of frequencies in the speech ('t Hart, Collier & Cohen, 1990: 75). A positive correlation exists between the two dimensions of pitch range; the higher the F0 level, the wider the pitch span (Urbani, 2012). In Gussenhoven's view (2004), the phonetic implementation of pitch values has generalizability in the paralinguistic intonational meanings and can be explained using a number of biological codes. For instance, the Frequency Code, which is based on the correlation between the size of the larynx and the rate of vocal fold vibration, suggested that a higher pitch frequently signals a smaller larynx and expresses speaker uncertainty, whilst a lower F0 tends to be associated with a larger organ of production and an assertive interpretation (Ohala, 1984, 1994) .

Based on the universality of biological codes, however, some languages were found to go against the general form-function patterns and show some language-specificities in the use of pitch codes. Evidence for this tendency of cross-language differences can be found in a growing number of recent studies in which different languages or language varieties have been reported to have different levels of pitch range and F0 variability in the speech. A crucial difference was found, for example, between tone languages such as Mandarin Chinese and stress languages like English. Compared to English, Mandarin Chinese shows generally higher pitch level, greater F0 fluctuation, and faster pitch change rate in the prose passage (Keating & Kuo, 2012; Eady, 1982) as well as wider F0 span in broadcast news speech (Yuan & Liberman, 2014). Similar F0 features of higher register and wider span were also reported in the Chinese dialect Min when compared with American English (Chen, 2005). Nevertheless, in Cantonese-English bilingual children's speech, significantly lower values of speaking fundamental frequency and F0 range were found for Cantonese speakers than for English-speaking children (Ng, Hsueh, Leung, & Shing, 2010). This difference in pitch implementation might be correlated with the tonal structure of Cantonese, although language proficiency and sociocultural conventions may also play a role in the

divergence of voice pitch characteristics. Furthermore, a recent study on cross-language comparison suggested a wider F0 range and higher F0 register in native Chinese speech than in native Japanese speech (Shi, Zhang, & Xie, 2014). However, when Japanese was compared to American English, Spanish and Dutch, it was found to have the highest values of mean F0 in read speech (Hanley, Snidecor, & Ringel, 1966; Van Bezooijen, 1995). The higher pitch values found for Japanese speakers, particularly for Japanese women, are not due to mere physiological or anatomical differences across gender and language communities (Rendall, Vokey, & Nemeth, 2007; Van Dommelen & Moxness, 1995), but are more closely linked to the social-cultural behaviors and relative powerlessness image of Japanese women (Van Bezooijen, 1995). Aside from the above-mentioned cases, further support for cross-language pitch value differences can be found in a large number of studies on distinct language pairs (see Mennen et al., 2014 and Ordin & Mennen., 2017 for a short review).

Compared to the accomplishments of cross-language research, few studies have shed light on L2 speech deviations from the target language. Because the acquisition of L2 intonation always entails some degree of interaction between the L1 and L2 systems, it is expected that the cross-language differences in the F0 register and span may also appear in the L2 production. Generally, it is suggested that L2 learners have a compressed F0 span and less variable pitch when compared to native speakers of the target language. For instance, Chen (1972) and Juffs (1990) reported that Chinese L2 learners had a narrower F0 span than native speakers of English. Recently, using a large-scale dataset extracted from a language learning app, Yuan et al. (2018) reconfirmed that, compared to that of native English speakers, the speech of Chinese L2 learners was characterized by a narrower pitch span, slower pitch change rate and more small "ripples" on the F0 contour. In addition, it was found that Chinese learners of L2 English and L2 German had higher values than native speakers in F0 span on the phoneme level, and in pitch change amount on the utterance level, due to the negative influence of L1 mandarin prosody (Ding, Hoffmann, & Hirst, 2016; Ding, Jokisch, & Hoffmann, 2012). This general trend of compressed range and less F0 variability in L2 speech can be observed in many L1-L2 combinations, for example, in Spanish learners of L2 English (Backman, 1979), in Chinese learners of L2 Japanese (Shi et al., 2014), among many others. The consistency seen in L2 pitch implementation patterns is probably influenced by the L1 prosody, but more frequently it has been correlated with learners' lack of confidence or cautiousness when speaking a non-native language (Mennen, 1998; Shi et al., 2014; Volín, Poesová & Weingartová, 2015). Another plausible reason for those L2 speech deviations could be that learners are too focused on the segmental pronunciation and stress emphasis, thus, there might be a lack of attention given to extending and varying the F0 pitch in a native-like way (Zimmerer et al., 2014).

In light of all these findings, it seems that there is a universal trend of pitch range compression when speaking an L2. However, in previous studies, most experiments were conducted with a small number of subjects. As a result, the conclusions drawn are potentially less convincing due to the high variability of the F0 range within speakers (Ladd, 1996). Another problem concerning previous studies is that in some cases the pitch range has been treated as a unitary concept without distinguishing the level from the span (Mennen et al., 2014), and have analyzed the data using different quantification measures. Most importantly of all, to date, none of the studies have explored the L2 pitch implementation characteristics of Spanish by L1 Chinese learners, thus, this research would be interesting evidence for the investigation of a general pitch compression pattern in the L2.

3. The present study

As can be seen in the literature review, prior studies in L2 prosody learning give inconsistent evidence for the cross-language differences in the phonological and phonetic dimensions, either due to the different speaking materials and quantification methods used in the research or because of the different language pairs under investigation. Of these studies, few have shed light on the production of Spanish intonation by L1 Chinese speakers, probably owing to the typologically substantial differences between the two language systems. Therefore, with the present study we intend to fill the existing gap in second language research and investigate the acquisition of intonation contours and pitch implementation details in L2 Spanish, by taking into account proficiency level (or L1), question type, gender and stress position. Specifically, the current study addresses the following questions:

1. Is the L1 prosodic system transferred (either positively or negatively) to the L2 Spanish intonation and therefore can the L1 account for some L2 deviation errors? If so, does the acquisition of L2 intonation patterns reflect different levels of proficiency?

2. Does the acquisition of pitch implementation details (as measured by six pitch variables based on the F0 distribution: mean F0, max F0, min F0, 100% span, 80% span and Pitch dynamism quotient -PDQ-) in an L2 reflect different levels of proficiency, and do they differ among different question types or stress positions as well as between male and female speakers?

3. Do our findings of L2 pitch range variation point towards a universal developmental trajectory (narrower span and less variable pitch) during the

L2 learning process, or are they highly dependent on the L1-L2 language pairs under study?

4. Does the acquisition of L2 Spanish intonation show different levels of difficulty depending on pragmatically different question types, and, if so, does this difficulty ranking exist only in the phonological dimension, or it can also appear in the phonetic dimension?

4. Methodology

4.1. Participants

The subjects of the present study were 5 native speakers of Peninsular Spanish and 32 learners of Spanish with Mandarin Chinese as their first language. The age of the participants ranged from 18 to 31 years (mean age: 23.97; SD=2.872). None of the individuals reported any speech, hearing or communicative impairments. The native control group consisted of 5 women who were born and/or lived for more than 20 years in Barcelona, and who had a comparable level of education (mean age: 23.2; SD=4.87). Although some of these participants were Catalan-Spanish bilinguals, they reported that Spanish was their dominant language.

As for the Chinese speakers (26 females and 6 males), they were all students and lived in Barcelona at the time of the recordings (mean age: 24.09; SD=2.53). Peninsular Spanish was the language variety (dialect) to which they had been predominantly exposed both during their leaning period in China and their immersion period in Spain. The L2 participants were divided into 2 groups according to proficiency level in line with the Common European Framework of Reference for Languages (CEFR): intermediate level (B1-B2) and advanced level (C1-C2). The Spanish language proficiency of the Chinese speakers was judged using the official language qualification DELE (Diploma of Spanish as a Foreign Language), with the exception of those learners who did not have this certificate. In the latter case (roughly 15% of the L2 learners), participants were required to state their self-evaluated L2 proficiency on the basis of Spanish language courses they had completed. In order to ensure that the learners were aware of the criteria of self-assessment, explicit descriptions of the six levels of European language proficiency were explained to those speakers at an early stage of this process.

In this study, we did not specifically control for Chinese learners' origin, age of L2 acquisition, or length of exposure to the target-language environment, due

to the dramatic reduction in subject pool which would result from including these selection criteria. However, as these variables were reported in previous literature to exert certain effects in the L2 speech (Cadierno et al., 2020; Juan-Garau & Pérez-Vidal, 2007; Kharkhurin, 2008; Pfenninger & Singleton, 2016), we decided to include this additional information for the non-native groups (see Appendix A for more details). The L2 participants were native speakers of Mandarin Chinese and declared that this was their dominant language, despite having different places of origin within China. The majority of the learners acquired Spanish in adulthood (mean age: 18.81; SD=2.08), only two female participants reported being teenage learners (they started to learn Spanish at 12 and 17 years old). Although the number of months of exposure to the target language was quite different between individual learners, the mean exposure time of advanced speakers (mean length:19.13; SD=9.51) was generally longer in comparison to the intermediate group (mean length: 22.8; SD=18.02).

4.2. Materials and recording procedures

In order to collect natural speech, we used the DCT (Discourse Completion Task) (Billmyer & Varghese, 2000; Félix-Brasdefer, 2010; Golato, 2003) method to elicit the corpus. Specifically, 15 daily scenarios were designed to elicit five question types with different linguistic meanings in the target language, namely, information-seeking *yes-no* questions ('YN'), information-seeking *wh*-questions ('WH'), disjunctive questions ('DJ'), confirmation-seeking *yes-no* questions ('CYN') and confirmation-seeking tag questions ('TAG'). Each question type varied in nuclear stress position (two positions: final and penultimate stressed syllable). Test items were mostly comprised of words with high familiarity ratings or high frequency (Tanaka & Terada, 2011), for the benefit of non-natives' comprehension during the task activity (see Appendix B for more details). The average syllable number in the study was 5.8 per utterance.

Situational contexts were presented by an interlocutor with whom the participants were somewhat familiar, and speakers were asked to produce the target sentence used in that situation. The task was performed only once except in cases where there was a problem with the speaker's first realization. Subjects were allowed to reproduce the test item if they made a mistake. All recordings took place in a soundproof room with a head-mounted microphone. Speech files were digitized at a sampling rate of 44.1 kHz and with a quantization precision of 16 bits. Each utterance was saved separately as a *wav* format file and annotated to a *TextGrid* object using a *Praat* script.

4.3. Data collection and analysis

4.3.1. Intonation contours

For the benefit of intonation labeling, unvoiced segments were interpolated through and F0 trace was smoothed using a *Praat* script with bandwidth set to 10 Hz. In our study, intonation transcriptions were realized by combining the visual representation of the F0 curve with the auditory perception of pitch accents. To facilitate the transcription work, all test items were initially annotated using a prosodic tool (Eti-ToBI) which automatically labels intonational events in Spanish utterances (Elvira-García et al., 2016). However, as this script was developed based on Sp_ToBI and Cat_ToBI conventions for native speakers, many unexpected pitch movements produced by Chinese learners could not be appropriately assigned (mainly due to differences in pitch alignment). Thus, manual correction of all labeling was conducted by the first author of the paper, according to the guidelines for Castilian Spanish intonation (Vilaplana, 2008; Vilaplana & Prieto, 2010). Furthermore, the annotation results were checked and revised by the second author of the paper and a third expert in Sp_ToBI labeling.

To conduct a cross-language comparison of intonation contours, the proportion of occurrence of pitch accents, boundary tones and nuclear configurations was calculated separately within the 5 question types. In this study, we were particularly interested in the nuclear configuration realization which has been referred to as the most salient part of an intonation contour (Prieto & Roseano, 2019) although it is worth noting that pitch variation is typically accompanied by changes in other prosodic features, like duration, amplitude, or voice quality. Regarding the set of linguistic functions that intonation (together with other prosodic factures, other parts of the F0 contour were excluded from further analysis.

4.3.2. F0 pitch extraction

Pitch tracking was performed automatically in *Praat* using the ESPS algorithm ('get F0') (Talkin, 1995), with pitch floor set to 70 Hz and pitch ceiling 600 Hz. A time step of 10 ms was used for the computation of F0. The original F0 data was refined by a manual correction based on *Pitch* objects. Specifically, cases with octave jumps and measurement errors (e.g. mistakes caused by creaky voice or laryngealizaiton, or false voicing in silent fragments) were marked as "unvoiced" and therefore excluded from further analysis. The speakers' pitch characteristics were analyzed along three dimensions: (a) *pitch level*, which has been referred to as a "reference line" (or a constant value) averaged by the rising and falling movements in the overall pitch contour

(Gussenhoven, 2004: 76); (b) *pitch span*, that is the distance between the highest and lowest values of the F0 contour, and finally, (c) *pitch variation*, which describes the degree of F0 variability in the speech (J't Hart, Collier & Cohen, 1990).

To quantify pitch characteristics in Spanish L1 and L2, two types of measures were included in the analysis. Firstly, for pitch level and span, we used the long-term distributional (LTD) measures developed by Mennen (2014). Specifically, 3 parameters have been extracted for level, and 2 parameters for span including both the absolute pitch excursion (100% span: max F0 - min F0) and the 80% span (the 90th and 10th percentile span) which has been reported to be the best F0 distributional measure (Mennen, Schaeffler, & Docherty, 2009; Niebuhr & Skarnitzl, 2019). Furthermore, because female and male speakers differed greatly in group size, we decided to introduce another pitch dynamism quotient measure (PDQ) to normalize the F0 variation data and to make the LTD results more robust. PDQ values were calculated by dividing the standard deviation by the mean frequency of each utterance. This metric gives an account of pitch variability and generally, the higher the PDQ, the greater the F0 variance and the emotional turmoil in subjects' speech (Shi, Zhang, & Xie, 2014; Wang & Qian, 2018; Zimmerer et al., 2014). To sum up, the following six dependent variables were extracted for the comparison of L1-L2 pitch implementation characteristics:

- Pitch level: *min F0*, *max F0* and *mean F0*
- Pitch span: *100% span*, *80% span*
- Pitch variability: *PDQ*

To assess the effect of scales, pitch measures of the span were also transformed into another two psycho-acoustic scales. Compared to the corresponding linear scale (Hz), the logarithmic (semitones) or the near-logarithmic (ERB-rate) scales have been reported to be the best measures for modeling intonational equivalence between females and males, and for capturing the frequency differences across speakers and languages (Nolan, 2003; Patterson & Ladd, 1999). Besides, correlation coefficients were calculated using the Pearson method in order to examine whether, and to what extent, F0 span is correlated with other pitch variables.

4.4. Statistical analysis

During the first stage, the between-group differences with regard to the production of pitch accents and boundary tones were tested in the R environment (R Core Team, 2016) using the chi-square statistic with continuity correction when all the cells of

the contingency tables for expected values had a value greater than 5. In the case that this condition was not met, we applied Fisher's exact probability test to examine the significance. At the second stage, statistical analysis of pitch results was performed using the linear mixed effects model. The model was run separately for six dependent pitch variables (namely, min F0, max F0, mean F0, 100% span, 80% span and PDQ) with Proficiency, Gender, Question type and Stress position, as well as all their possible interactions as fixed effects. Subjects were included as random effects with all possible random intercepts. We used the Anova function to test the significance of main effects, and *p* values were fitted by eliminating non-significant effects of the initial model with the *lmerTest* package and evaluated using Satterthwaite's approximation (Kuznetsova, Brockhoff, & Christensen, 2017). The post-hoc comparisons were conducted using the single-step function of the *multcomp* package (Hothorn et al., 2016) supported by the *emmeans* function (Lenth, Singmann, & Love, 2019). Finally, the correlation coefficients of various F0 distributional measures were obtained using Pearson's method (Benesty et al., 2009).

5. Results

5.1. Results of overall pitch accents and boundary tones

The first analysis examines the proportion of pitch accents and boundary tones of the five question types within each of the three language groups, namely, L2 Chinese intermediate learners (hereafter CI), L2 Chinese advanced learners (hereafter CA), and L1 Native Spanish speakers (hereafter SN). Of the 555 speech items, 119 items were realized with a low-rising accent L+H* and 61 items, with a high-level tone H* in the nuclear location, regardless of final pitch movements. Interestingly, both accent types were produced by Chinese L2 speakers, with a relatively higher proportion for the CI group than for the CA group (see Table 1). However, the probability test revealed no statistical significance between the two learner groups with regard to the production of the two pitch accents H* (χ^2=1.20, p>0.1) and L+H* (χ^2=0.72, p>0.1). The vast number of high plateaux and rising patterns in L2 Spanish corroborate previous findings for L2 English and L2 German spoken by Chinese learners (Ding et al., 2012; Ji et al., 2009, 2012). Nevertheless, concerning the nuclear accent L*, the SN group used a significantly higher number of steady low tones compared to the CI (χ^2=8.94, p<0.01) and CA learners (χ^2=5.12, p<0.05). Similar results have also been observed in the realization of the pitch accent H+L*, whereby Spanish L1 speakers presented a significantly higher proportion of falling contours than the two learner groups (CI-SN: p<0.001; CA-SN: p<0.01).

Table 1 Proportion of pitch accents produced by the three language groups over the five question types

	H*	L+H*	L+¡H*	L*	H+L*	Total
CI	14.58%	26.67%	0%	58.75%	0%	240
CA	10.83%	22.92%	0%	63.75%	2.50%	240
SN	0%	0%	10.67%	78.67%	10.67%	75

In regard to the boundary tones, Table 2 shows that, compared to the Spanish native group, Chinese L2 learners, particularly the CI group, produced significantly more rising boundaries (H%: χ^2=4.39, p<0.05) and less falling boundaries (L%: χ^2=5.99, p<0.05) over the five question types. In addition, a small proportion of rising-falling boundaries (HL%) was exclusively observed in the L2 intonation (see Table 2). These results appear to be contradictory, to some extent, to our initial expectation that L2 learners tend to use a final fall more frequently than a final rise for Spanish questions, given that in their L1 Chinese questions are marked mostly by interrogative particles. A possible explanation for this phenomenon may be that Chinese learners have overproduced the typical final rise for all question types, due to a lack of knowledge about the intonational phonology of the target language (both its possible patterns and their communicative functions). Furthermore, our study seems to suggest a general trend of improvement during the L2 acquisition of phonological tones. More precisely, it was found that compared to Chinese intermediate learners, the advanced group showed systematically a more native-like performance in producing the target pitch accents and boundary tones. However, these between-group differences are not statistically significant (p>0.1 in all cases except for the nuclear accent H+L* where the p value was less than 0.05). This is likely because the oral proficiency levels of Chinese intermediate (B2 level) and advanced learners (mostly C1 level) were very similar in our research.

Table 2. Proportion of boundary tones produced by the three language groups over the question types

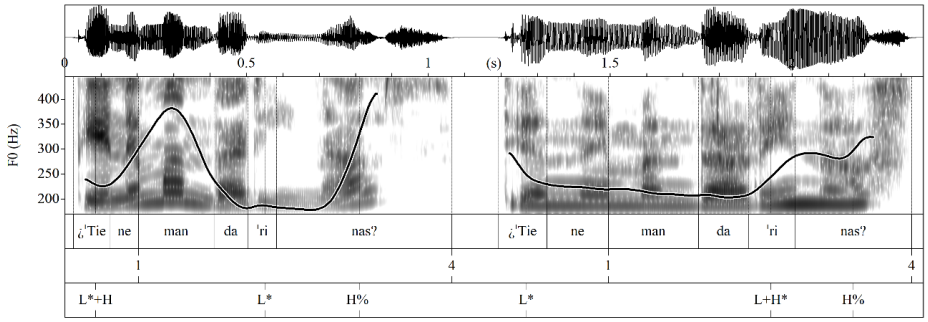
	H%	L%	HL%	Total
CI	67.50%	30.42%	2.08%	240
CA	64.17%	34.17%	1.67%	240
SN	53.33%	46.67%	0%	75

5.2. Results of nuclear configurations

5.2.1. Information-seeking yes-no questions

Since pitch contours did not differ significantly between the two Chinese groups (see the statistical results in section 5.1), we treated the non-native subjects as a single group when comparing them to L1 native speakers in the nuclear configuration analysis. Specifically, for information-seeking *yes-no* questions, only around 40% of the Chinese learners successfully acquired the native-like pattern (L* H%), which was significantly less than the Spanish L1 group who consistently used the typical low-rising contours for *yes-no* questions (χ^2=16.64, p<0.001). The majority of L2 speakers were found to apply an early rising accent (L+H*: 35%) or use high-level accents (H*: 18%) with a rising end (H%) on the nuclear position (see Figure 2). The large number of deviated accents produced by Chinese learners may be explained as a negative transfer from their L1. Mandarin Chinese does not have a steady low tone (L*) in its tonal inventory and intonational stress is realized mainly through the pitch range expansion of high tone, therefore, Chinese learners may unconsciously transfer the high-level tone (namely Tone1) or the rising tone (namely Tone2) from their L1 into the L2 Spanish prosody. Moreover, it was noted that most of the Chinese speakers were capable of producing the final rising boundaries (H%) of *yes-no* questions, with only 7% of the L2 speakers failing to achieve the high pitch targets in sentence-final locations.

Figure 2. Waveform, spectrogram and F0 trace of the information-seeking *yes-no* question "¿Tiene mandarinas?" '*Do you have Tangerines*?' produced by a native speaker (L* H%) (left) and a Chinese learner (L+H* H%) (right)

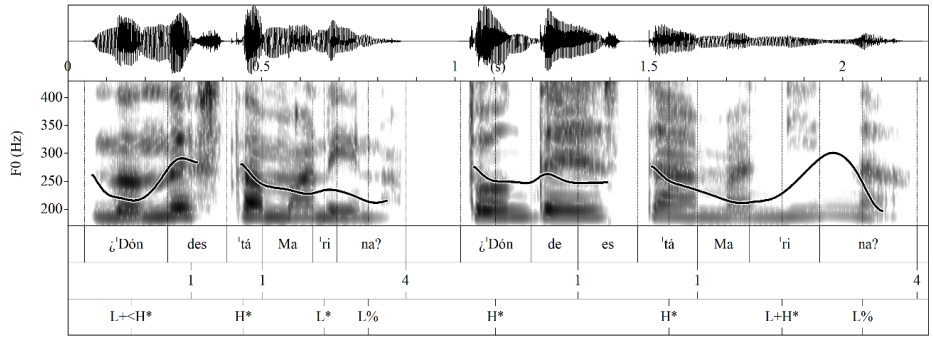


5.2.2 Information-seeking wh-questions

With regard to information-seeking *wh*-questions, the Spanish L1 speakers produced three possible pitch contours in the nuclear position: L* L% (30%), L* H% (18%) and L+¡H* L% (53%). Of these patterns, the first falling pattern (L* L%) has been interpreted as the most neutral and default realization of *wh*-questions, while the second rising contour (L* H%) expresses a nuance of politeness and the speaker's interest or greater involvement in the informative speech act (Estebas-Vilaplana & Prieto, 2010: 35). The third pattern L+¡H* L% is not an information-seeking *wh*-question, but has instead been described as a form of the reintroduction of a previously stated topic. The high proportion of this complex contour in our corpus can likely be attributed to different interpretations of the situational contexts by the L1 speakers.

As for the L2 intonation results, the Chinese learners showed a great probability of success in producing the two typical falling (L* L%: 30%) and rising contours (L* H%: 18%) used for the target *wh*-questions, and no statistically significant difference was found between the Spanish L1 and L2 groups (p>0.1). Aside from the native-like production, the rest of the L2 speakers were found to show deviation problems similar to those of information-seeking *yes-no* questions, that is, a tendency to apply rising tones (L+H*: 44%) or high-level (H*: 8%) tones instead of low accents (L*) in the nuclear location (see Figure 3).

Figure 3: Waveform, spectrogram and F0 trace of the information-seeking *wh*-question "¿Dónde está Marina?" '*Where is Marina?*' produced by a native speaker (L* L%) (left) and a Chinese learner (L+H* L%) (right)

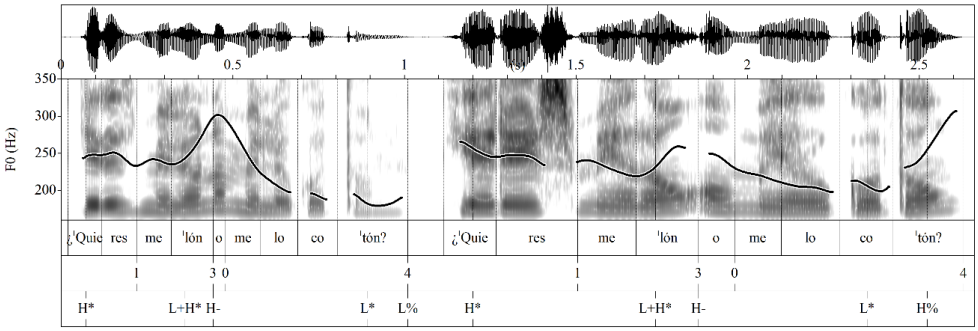


5.2.3 Information-seeking disjunctive questions

Disjunctive questions in L1 Peninsular Spanish were realized mainly with a rising contour (L+H* H-: 93%) in the first prosodic unit and a falling movement (L* L%: 100%) at the end of the utterance (see Figure 4). Among Chinese learners, in the

first prosodic unit, 81% of the participants successfully produced the target rising pitch pattern (L+H* H-), statistically they showed no significant difference compared to the L1 native performance (χ^2= 0.62, p>0.1). For the second prosodic group, more than half of the L2 learners (roughly 66%) were able to produce the final falling pitch contours (L* L%) found in disjunctive questions. Nevertheless, according to Fisher's exact test, there is a significant difference (p<0.01) between the L1 and L2 speakers with regard to their intonational performance. The high level of accessibility of target intonation contours of disjunctive questions by Chinese speakers could be explained as a positive transfer of L1 intonation strategies. In Mandarin Chinese, disjunctive questions are realized in a similar way to those of Peninsular Spanish, that is, by the expansion of pitch range in the first prosodic unit and compression in the last constituent of the sentence. This similarity in the phonetic dimension between the source and target languages appears to benefit or accelerate learners' rate of acquisition of target intonation patterns. In addition to the low tone (L*) deviations, it is also interesting to note that some Chinese learners might make mistakes when producing the intermediate and final boundaries of disjunctive questions. For instance, roughly 24% of Chinese speakers were found to use a high-rising boundary (H%) instead of a low-falling boundary (L%) in the final constituent of disjunctive questions (see Figure 4). This may happen because of their lack of intonational knowledge and the cognitive bias in marking questions with final rising pitch movements.

Figure 4. Waveform, spectrogram and F0 trace of the disjunctive question "¿Quieres melón o melocotón?" '*Do you want melon or peach?*' by a native speaker (L+H* H- L* L%) (left) and a Chinese learner (L+H* H- L* H%) (right)



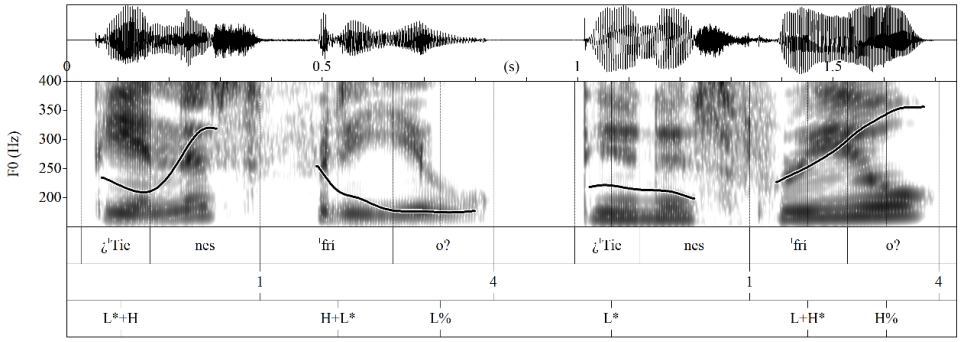
5.2.4 Confirmation-seeking yes-no questions

In addition to the prototypical informative speech act, questions in Peninsular Spanish can also be used to convey the illocutionary force of confirmation. Unlike information-seeking questions in which speakers do not have any expectations

regarding the answer, confirmation-seeking questions are biased questions where the information is previously given or shared through the context or presumptions, and the speaker usually expects a positive answer from the other interlocutor (Frota & Prieto, 2015; Vanrell et al., 2010). In our corpus, more than half of the L1 speakers applied a falling contour (H+L* L%: 53%) to express a relatively higher certainty on the proposition, whilst the rest (47%) maintained the canonical rising pitch pattern (L* H%) used for information-seeking *yes-no* questions.

As for the Chinese learners, only a small number (6%) had acquired the native-like falling pitch accent (H+L*) combined with a low boundary tone (L%) for this type of question. By contrast, 35% of the L2 speakers borrowed the nuclear contours of information-seeking *yes-no* questions, while the rest exhibited several different patterns which are congruent with a deviated form of L* H% with the low nuclear accent (L*) misproduced (see Figure 5). These results proved significantly different (χ^2=16.01, p<0.001) from those produced by L1 native speakers.

Figure 5. Waveform, spectrogram and F0 trace of the confirmation-seeking *yes-no* question "¿Tienes frío?" '*Are you cold?*' produced by a native speaker (H+L* L%) (left) and a Chinese learner (L+H* H%) (right)

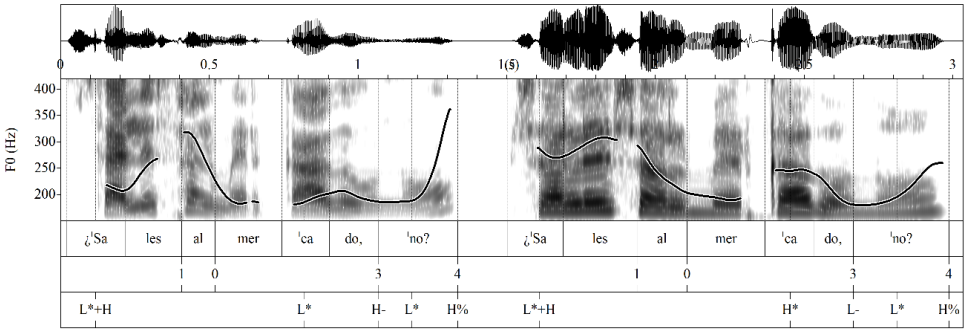


5.2.5 Confirmation-seeking tag questions

Confirmation-seeking questions can be produced, in addition to the typical falling pattern presented in the section above, with various confirmative tags (i.e. *¿no?* (no?), *¿verdad?* (is it?), *¿no es cierto?* (isn't it?)). This syntactic resource is valid both in Mandarin Chinese and in Peninsular Spanish, as it expresses a stronger commitment by the subject to the truth value of the proposition. In our experiment, all question tags were consistently produced with a low-rising pitch contour (L* H%) by Spanish L1 and L2 speakers. A possible explanation for this phenomenon has been correlated with the (alleged) universality in marking question tags with various alternative strategies

such as final rise, final pitch range expansion, late prominence and some emphasis gestures (Cuenca, 1997; Gussenhoven & Chen, 2000). Besides, the simple structure (mostly 1 or 2 syllables) of those tags has also been proposed as a plausible factor which accounts for the high intonational accuracy in L2 tag questions. Note, however, that the production of target phonological patterns does not necessarily mean that L2 speakers have acquired a 100% native-like prosodic performance. It has been observed, for example, that the question tags produced by Chinese learners differed phonetically from those produced by the L1 native speakers in the F0 differences of the final rising movements (see Figure 6). Nevertheless, this comparison is outside the scope of this paper.

Figure 6: Waveform, spectrogram and F0 trace of the confirmation-seeking tag question "¿Sales al mercado, no?" '*You are going to the market, aren't you?*' produced by a native speaker (L* H%) (left) and a Chinese learner (L* H%) (right)



5.3. Pitch implementation results

5.3.1. Pitch level results

This section aims to show the differences in pitch level (min, max and mean F0) between Spanish L1 and L2 speakers. The statistical analysis revealed a tendency among female learners of the CI and CA groups to have systematically higher pitch levels than female speakers of the SN group (see Table 3), due to the influence of L1 tonal structure. Nevertheless, results on main effects (Table 4) show no significance of Proficiency on any of the three measures of pitch level. On the contrary, the effects of Gender, Question as well as the interaction of Proficiency and Question on the three pitch measures. Furthermore, Table 4 shows that there is a mild effect of Stress on the mean F0 and the factor Stress strongly interacted with Proficiency on the variable of mean F0. No significant interactions were found between Gender and Question, Gender and Stress on the three measures of pitch level.

Table 3. Descriptive statistics of pitch level results depending on language group and gender

	Mean F0	**Min F0**	**Max F0**
CI_M	143 Hz	108 Hz	214 Hz
CI_F	236 Hz	189 Hz	329 Hz
CA_F	232 Hz	175 Hz	335 Hz
SN_F	220 Hz	167 Hz	327 Hz

Table 4. Effects (F values) of Proficiency, Gender, Question type, Stress position and its interactions on the three measures of pitch level (Hz) (*N=555, '***'p<0.001; '**'p<0.01; '*'p<0.05; ' •' p< 0.1*)

	Mean F0	**Min F0**	**Max F0**
Proficiency	1.06	2.50•	0.17
Gender	95.01***	63.17***	37.90***
Question	8.49***	5.73***	13.38***
Stress	5.80*	0.37	3.70•
Proficiency *Question	7.80***	3.65***	6.48***
Proficiency*Stress	8.12***	0.45	0.90

On the one hand, the lack of a significant effect of Proficiency, but its strong interaction with Question type on the three F0 measures, further suggests that pitch level may differ within each of the five question types produced by learners of different levels of proficiency. Specifically, the post-hoc analysis indicates that the CI group has a significantly higher minimum ($t(2)=2.59$, $p<0.05$) and mean F0 values ($t(2)=3.09$, $p<0.01$) in CYN questions than the L1 native speakers. Also, Figure 7 shows that speakers in the CI group tend to use a higher mean pitch in WH questions than the native participants ($t(2)=2.56$, $p<0.05$). By contrast, the SN group was found to have a significantly higher maximum F0 in WH questions when compared to the CA group ($t(2)=2.47$, $p<0.05$). On the other hand, Figure 8 indicates that the two Chinese groups tend to have significantly lower mean F0 (CI: $t(2)=-3.383$, $p<0.01$; CA: $t(2)=-4.359$, $p<0.001$) in utterances with final-syllable stressed words compared to those with stress on the penultimate syllable. Moreover, the CI group was found to use a statistically lower maximum F0 in final-syllable stressed questions ($t=-2.45$, $p<0.05$). These results differ from those of the SN group, which exhibited higher pitch values in final-syllable stressed sentences (oxytone words), as described in Figure 8. A possible explanation for

this deviation in L2 pitch has been correlated with the extra difficulty of combining intonational patterns and those required by lexical stress, since some pitch contours require adaptations when they are applied to oxytone words (e.g. L* H% becomes L+H* L%) (Prieto & Roseano, 2019)although it is worth noting that pitch variation is typically accompanied by changes in other prosodic features, like duration, amplitude, or voice quality. Regarding the set of linguistic functions that intonation (together with other prosodic factures.

Figure 7: Mean pitch of the three language groups depending on proficiency and question type. Upper and lower levels indicate maximum and minimum F0

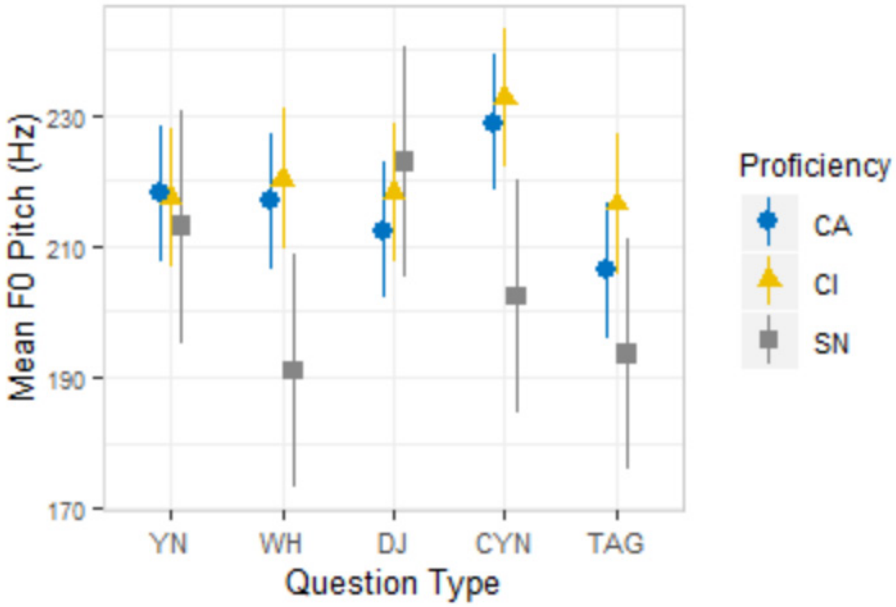
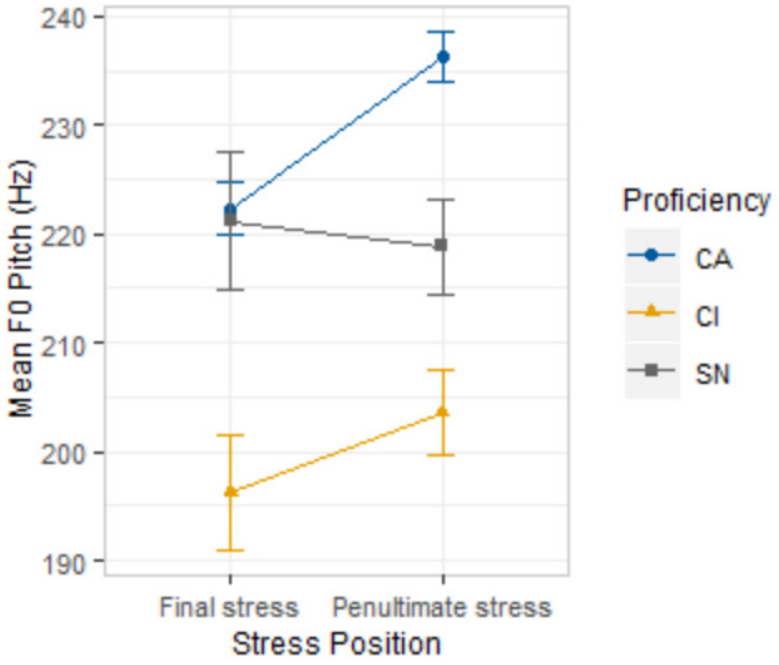


Figure 8: Mean pitch of the three language groups depending on proficiency and stress position. Error bars indicate ±1SE



5.3.2. Pitch span results

Of the two F0 span measures, only the 80-percentile span measured in St showed a significant main effect of Proficiency ($F(2, 38)=3.47$, $p<0.05$). As expected, the two learner groups showed a more reduced span than the L1 native speakers (80% span: CI:7.11 St; CA: 8.34 St; SN: 9.45 St). Nevertheless, the F0 differences were only statistically significant when comparing the CI and SN groups ($t(2)=-2.53$, $p<0.04$). Besides this, there was a strong main effect of Question type (80% span: $F(4, 518)=9.16$, $p<0.001$; 100% span: $F(4, 518)=12.30$, $p<0.001$) as well as its interaction with Proficiency (80% span: $F(8, 518)=8.78$, $p<0.001$; 100% span: $F(8, 518)=7.01$, $p<0.001$) on the two span measures. Since statistical results among scales of Hz, St and ERB were similar on Question type and on the interaction between Proficiency and Question type, only the St version is reported in this section.

Furthermore, a post-hoc test which looked for the interaction between Proficiency and Question on the 80% span showed that within the five question types, the CI and CA groups tended to use a significantly narrower span than the L1 speakers in YN questions (CI-SN: $t(2)=-3.38$, $p<0.01$; CA-SN: $t(2)=-2.82$, $p<0.05$), DJ questions (CI-SN:

t(2)=-3.70, p<0.01; CA-SN: t(2)=-3.53, p<0.01) and CYN questions (CI-SN: t(2)=-2.53, p<0.05). By contrast, regarding WH questions, Figure 9 indicates that the two Chinese groups had a wider span than the native speakers, although this trend did not reach statistical difference in our study. Besides, the factor Proficiency was also found to interact significantly with Stress on the 80% span measured in ERB (F(2, 518)=3.54, p<0.05) and in Hz (F(2, 518)=4.33, p<0.05). More precisely, the post-hoc tests on all scales indicated that the two learner groups, particularly the CA group (e.g. St: t(2)=-2.70, p<0.01), tended to compress the span in questions with final-syllable stressed words more than those with stress on the penultimate syllable. By contrast, results in Figure 10 seem to suggest an opposite trend for the SN group regarding the span performance on the two stress positions. Finally, unlike the mean F0, no significant main effect of Gender or Stress was found on the span measures on any of the scales.

Figure 9. 80% F0 span of the three language groups depending on proficiency and question type

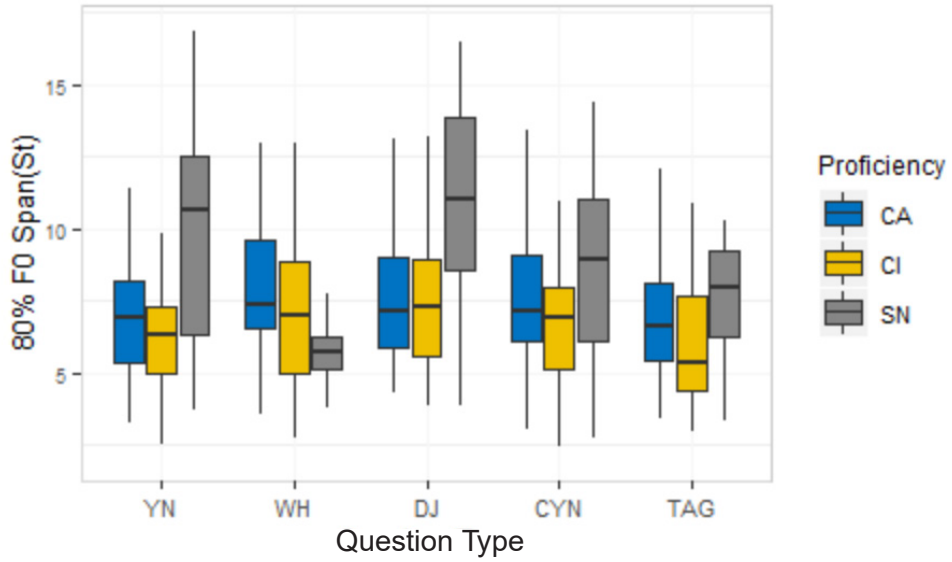
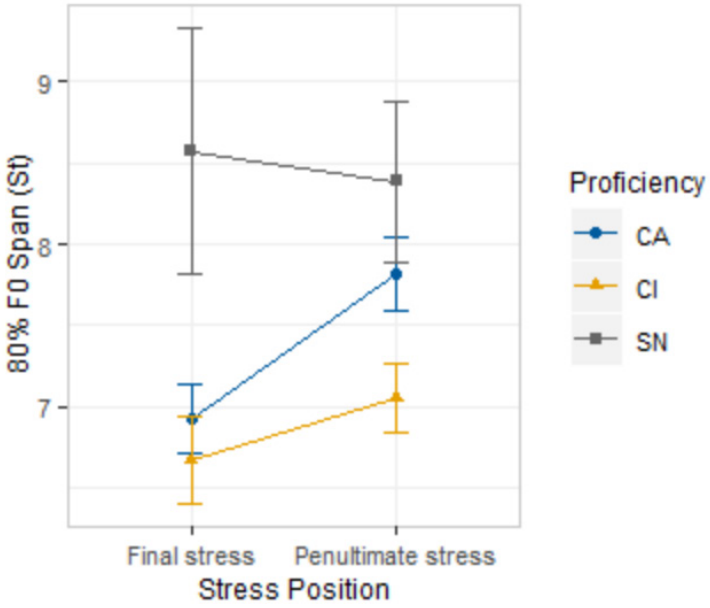


Figure 10. 80% F0 span of the three language groups depending on proficiency and stress position. Error bars indicate ±1SE



5.3.3. Pitch variation results

With regard to the pitch variation, the statistical analysis showed a significant main effect of Gender ($F(1, 38)=4.56$, $p<0.05$), Question type ($F(4, 518)=10.07$, $p<0.001$) and Stress position ($F(1, 518)=3.98$, $p<0.05$). To be specific, our results showed that female speakers were significantly more variable than male speakers in their use of F0, and as a whole the participants had higher F0 variance (or higher pitch dynamism quotient -PDQ-) in DJ questions (mean: 0.214), YN questions (mean: 0.211) and CYN questions (mean: 0.200) than in WH (mean: 0.173) and TAG questions (mean: 0.171). In addition, Figure 12 indicates that speakers of all three proficiency levels had consistently higher F0 variability in questions with penultimate syllable stressed words (mean: 0.200) than those with stress on the final syllable (mean: 0.188). In addition to those features, it is interesting to note that the two Chinese groups produced an overall lower PDQ (CI: 0.168; CA: 0.198) than the L1 native speakers (SN: 0.216), although the factor Proficiency did not show significant main effect in this analysis. Nevertheless, the statistical model revealed a strong interaction between Question type and Proficiency ($F(8, 518)=8.30$, $p<0.001$). As can be seen in Figure 11, compared to the two Chinese groups, the Spanish L1 speakers were higher in PDQ for all question types, in contrast to WH questions in which native speakers produced the lowest PDQ. More precisely,

the post-hoc test indicated that the native group had statistically higher PDQ in DJ questions than the CI ($t(2)=-3.317$, $p<0.01$) and CA groups ($t(2)=-2.593$, $p<0.05$), as well as in YN questions when compared to the CI group ($t(2)=-3.218$, $p<0.01$). No other factors or interactions reached significance on the variable of PDQ.

Finally, in order to examine whether, and to what degree, the F0 span is interdependent on other pitch variables (i.e. mean F0, max F0, min F0 and PDQ), we analyzed the correlation between the span and the rest of the variables that we have taken into account in the pitch range analysis. The statistical results indicated that the 100% and the 80% span were more closely correlated with the maximum F0 (100% span: $r=0.84$, $p<0.001$; 80% span: $r=0.66$, $p<0.001$) than with the mean F0 (100% span: $r=0.38$, $p<0.001$; 80% span: $r=0.41$, $p<0.001$) or the minimum F0. These results appear to suggest that the more a speaker is able to approximate to the high pitch targets, the wider the entire F0 range of his/her speech. Additionally, there was a strong positive correlation of PDQ with the 100% span ($r=0.78$, $p<0.001$) and the 80% span ($r=0.80$, $p<0.001$), as illustrated in the right panel of Figure 13. The high consistency between the two span measures and the PDQ measure also consolidated the results of pitch range variation in our study, despite the unbalanced group size across gender and language group. In general, our findings seem to suggest that the wider the F0 span, the more variable the speech.

Figure 11. Mean PDQ of the three language groups depending on proficiency and question type

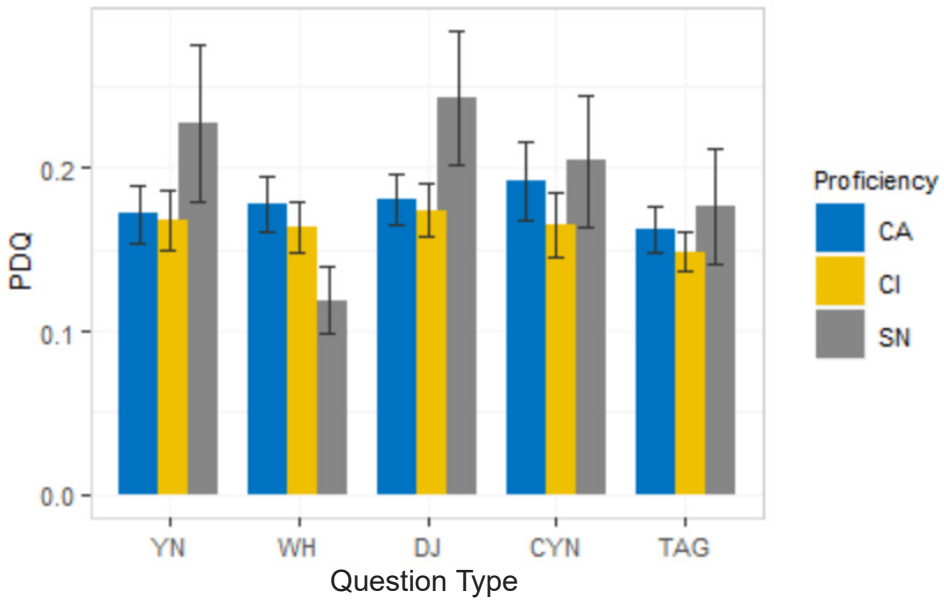


Figure 12. Mean PDQ of the language three groups depending on proficiency and stress position

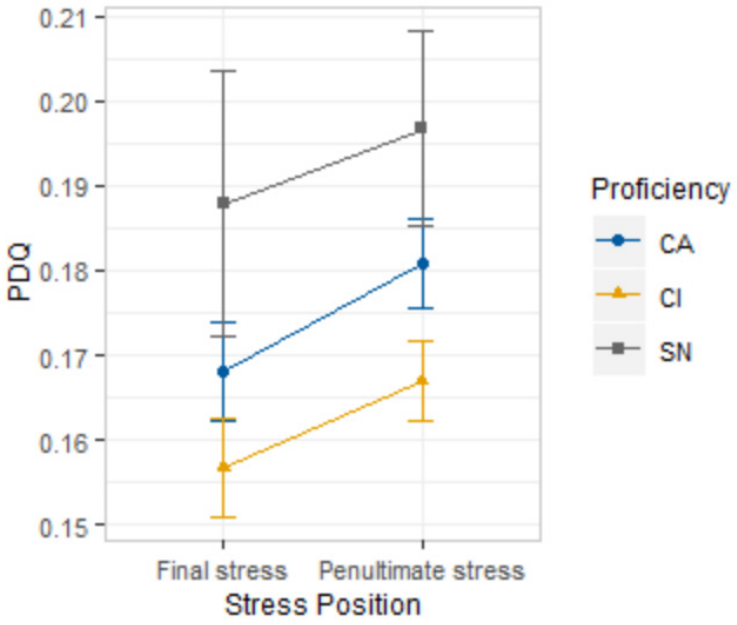
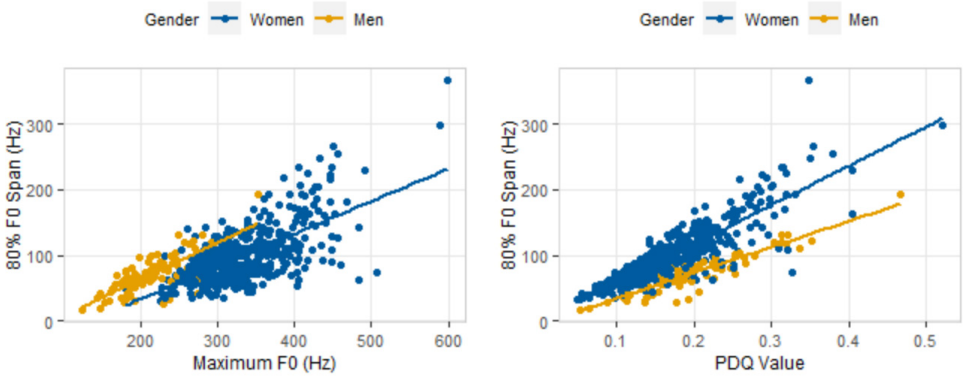


Figure 13. Correlation of the 80% F0 span with the maximum F0 (left) and with the PDQ (right)



6. Discussion

The present study sets out to investigate the acquisition of L2 Spanish prosody by Chinese L1 speakers. For this purpose, percentages of occurrence of the pitch accents,

boundary tones and nuclear configurations related to different speech acts, as well as various pitch implementation variables have been examined across the three language groups.

The first research question enquired whether the acquisition of intonation contours in an L2 reflects different levels of proficiency, and whether prosodic knowledge from the L1 can be transferred onto L2 pitch movements. In general, it was found that the L2 intonation performance of the phonological contours is dependent on proficiency level, and most of the intonation deviations and success in the L2 could be accounted for under an L1 transfer hypothesis; therefore the answer to the first research question is affirmative. Specifically, the descriptive statistics suggested that compared to the L2 intermediate speakers, L2 advanced learners generally acquired more native-like patterns in producing pitch accents and boundary tones. However, these differences between the two learner groups did not reach statistical significance in our study, probably due to the modest gap of oral proficiency between the Chinese intermediate (B2 level) and advanced group (mostly C1 level). Furthermore, our study found that both the L2 intermediate and the L2 advanced groups produced a significantly higher proportion of high plateaux (H*) and rising accents (L+H*) in the nuclear location than the Spanish L1 speakers, who did not produce any of the aforementioned accent types but showed a statistically higher number of low tones (L*) or falling movements (H+L*) instead. These deviation errors produced by Chinese speakers have been mostly correlated with the negative transfer of their L1. As described before, the lack of a steady low tone in Mandarin's lexical tone inventory and its different stress realization mechanisms (through the expansion of high tone) may be contributing factors in the frequent use of high/rising accents in L2 Spanish intonation. Generally, our findings regarding high/rising tone strategies are consistent with previous studies in L2 speech, which also reported a similar tendency for Chinese learners of other intonation languages such as English and German (Ding et al., 2012; Ji et al., 2009, 2012; Yuan et al., 2019). Moreover, it was surprising that the two Chinese groups, particularly the intermediate group, produced a significantly larger number of high boundaries (H%) and fewer low prosodic tones (L%) than L1 native speakers at the end of the five question types. Such results were contrary to our initial expectations that Chinese learners would use more final falls than rises in L2 Spanish intonation, due to the negative influence of L1 question marking strategies using a particle. A possible explanation for this phenomenon might be that Chinese speakers had at their disposal few intonation contours of the target language and therefore could not assign the correct pattern of different sentence types with specific functional meanings, thus, they tended to overproduce all the questions with the most prototypical final rising intonation. Nevertheless, in general, our results seem to suggest that the performance of L2 speakers in the production of pitch accents and boundary tones can progress

towards native-like shapes by increasing their proficiency level and phonological awareness in the L2.

As for the nuclear configurations, it was found that, excluding the low-rising contours (L* H%), more than half of the learners deviated from the L1 native speakers in the overproduction of high *plateaux* (H* H%) or early rising accents (L+H* H%) in information-seeking *yes-no* questions. Moreover, unlike Spanish L1 speakers who consistently used a high boundary tone to mark the information-seeking *yes-no* question, the final boundaries of Chinese learners could be either high (H%) or low (L%), although low boundaries constituted only a small proportion in their L2 intonation. These results corroborate previous studies on L2 intonation (Ding et al., 2012; Ji et al., 2009, 2012) that have reported similar problems for Chinese L1 speakers of L2 English regarding *yes-no* question patterns. The prevalence of high/rising accents was also observed in the nuclear accents of information-seeking *wh*-questions and confirmation-seeking *yes-no* questions, as a result of the systematic negative transfer from L1 prosody explained above. But in the case of confirmation-seeking *yes-no* questions, most L2 learners have borrowed the intonation inventory of information-seeking *yes-no* questions and only a few of them were able to produce the most salient falling pattern (H+L* L%) for the confirmative speech act. The high comparability in the realization of the two questions types might be attributed to the fact that they share the same syntactic structure in Spanish. However, it is possible that learners faced difficulties in correlating different functional meanings with the specific pitch contour shapes in the L2. Besides, we found that the nuclear patterns of *wh*-questions were variable even in L1 native speech, likely due to individual differences in the interpretation of the situational contexts. Whilst the falling (L* L%) and rising contours (L* H%, with a nuance of politeness and interest) tend to be interpreted as neutral contours for *wh*-questions, the complex rising-falling patterns (L+¡H* L%) in our dataset were biased forms that focused on the reintroduction of topics instead of requesting information. Overall, it seems that the rate of acquisition of target intonation contours is dependent, to some extent, on the systematic similarities between the first prosody and the target prosody, and in this way, we can reasonably explain the great success of Chinese learners in producing L2 disjunctive questions and tag questions as a positive transfer of L1 intonation.

The second research question asked whether the L2 acquisition of pitch implementation details reflects different levels of proficiency, and whether the pitch range values differ among question type, stress position, and gender. The overall results seem to give an affirmative answer to the first part of the question. More precisely, our study found a better pitch performance for the L2 advanced learners than for the L2 intermediate speakers, although the factor "proficiency" did not reach

significance on any of the three pitch measures (except the 80% span measured in St). In addition, the significant interactions of proficiency with other dependent variables show that the L2 prosody acquisition is more complex than previously stated and the proficiency effect was modulated by question type and stress position. Furthermore, the significant main effect of question type for all the pitch measures appears to give a positive answer to the second research question, showing that different sentence types are encoded with different pitch profiles which serve different pragmatic purposes. In other words, the language-specific pitch implementation values were highly dependent on question type. Specifically, we found that the intermediate learners' pitch increased significantly in information-seeking *wh*-questions and confirmation-seeking *yes-no* questions. In addition, it was observed that the two learner groups compressed the span and the F0 variability in all of the question types except for the *wh*-questions in which Chinese informants used wider span and greater F0 variation compared to the L1 native speakers.

As for the pitch differences on stress position, it was interesting to observe that the Chinese learners tended to increase the F0 register, expand the pitch span and exhibit more frequent F0 contour variation in questions where the stress falls on the penultimate-syllable of the word than those with stress on the final syllable of the word, whereas the native speakers (on the contrary) produced a wider span and higher register for the oxytone words. The factor which accounts for these differences between L1 and L2 is mainly correlated with learners' unfamiliarity and lack of knowledge of rules in producing the final-syllable stressed words, particularly when they appear in the sentence-final position that requires compressed pitch movements. In such circumstances, learners may pay more attention to the pronunciation of words and reduce the use of F0 strategies in the L2 (Mennen, Schaeffler, & Dickie, 2014; Peters, 2019; Yuan et al., 2018; Zimmerer et al., 2014). Finally, with regard to gender effects, our results are congruent with previous findings which report that women use significantly higher pitch level and greater pitch variation than men. Hence, the answer to the latter part of the second research question is also affirmative. According to Ohala (1994), Van Bezooijen (1995), Urbani (2012), Mennen et al. (2014) and Peters (2019)weak, dependent, modest, pitch differences across gender are not merely driven by physiological and anatomical factors, but are also closely linked to some sociocultural aspects such as social status and gender roles. Moreover, it is assumed that the greater F0 variation of female speakers found in our study might be correlated with the finding that women tend to express many emotions more frequently than men, except for pride and power (Brebner, 2003). However, further research on the current topic are required to provide direct and strong evidence for our speculation. Finally, in the data obtained for our study, we also found a strong positive correlation between the F0 span and the maximum F0 as well as the F0 variability. More precisely,

our study appears to suggest that the wider the entire F0 span, the higher the F0 pitch targets and the more variable the F0 speech.

The third research question enquired whether our pitch implementation findings point towards a universal developmental path during the L2 learning process, regardless of the language combinations under study. In the present study, we found that compared to Spanish L1 speakers, Chinese learners tended to use higher pitch level (specifically higher minimum and mean F0), narrower span and less variable pitch in the L2. These results are in agreement with other preliminary studies on L2 speech which also documented a compressed pitch pattern for L2 speakers of different language backgrounds, and reported a similar trend towards improvement as their L2 experience increased (Busà & Urbani, 2011; Mennen, Schaeffler, & Docherty, 2009; Mennen, 1998; Peters, 2019; Shi, Zhang, & Xie, 2014; Ullakonoja, 2007; Urbani, 2012; Yuan et al., 2018; Zimmerer et al., 2014). As for the differences of F0 height, the higher pitch level observed in L2 intonation has been explained as a result of an increased cognitive effort when speaking a non-dominant language (Zimmerer et al., 2014). However, in our study, we speculate that the L1 pitch characteristics may also play a role in the F0 rise of L2. To test this hypothesis, further investigations comparing L1 Chinese and L1 Spanish are needed.

The final research question examined whether the acquisition of the phonological and phonetic patterns of an L2 reflects different levels of difficulty corresponding to pragmatically different question types. On the one hand, it is interesting to note that the probability of success with regard to the production of L2 nuclear contours was different across the five question types. Therefore, the answer to the last research question must be affirmative in the phonological dimension. Specifically, our results seem to suggest a hierarchy of difficulty in the intonation learning of different question patterns of Spanish by Chinese L1 speakers, whereby the confirmation-seeking *yes-no* question was the most difficult pattern, followed by the information-seeking *yes-no/wh* question, then the disjunctive question and finally the confirmation-seeking tag question. Along the same lines, Cortés Moreno (1999, 2004), Liu (2005), Mennen (2015) and Yuan et al. (2019) have also reported that different sentence types may imply different degrees of difficulty during the learning process of L2 phonological contours. On the other hand, the varied pitch performance among the five question types appears to extend previous hypotheses on L2 intonation (Cortés Moreno, 1999, 2004; Liu, 2005; Yuan et al., 2019) from the phonological to the phonetic dimension, suggesting that there is also a ranking of difficulty in implementing the pitch values of different question patterns in L2. The relatively faster rate of acquisition of the two intonational aspects (phonological and phonetic) has been partially correlated with the typological closeness between the first language and the target language.

Moreover, this has also been explained as a result of the "perceptual salience" of some intonational movements in the target language (Yuan et al., 2019).

7. Conclusion

All things considered, the present study has shown that the L2 intonation learning process is more complex than previously stated, whereby the phonological and phonetic dimensions may develop in a non-parallel way. Specifically, it is worth noting that learners who have successfully acquired the target-like intonation contours may still deviate from native speakers in the pitch implementation details of the L2 prosodic system. However, as their L2 proficiency increases, they are capable of approximating to the target language settings both in the systematic dimension and in some realizational aspects. This finding of non-uniform development paths for L2 phonetic and phonological acquisition has also been observed in other L1-L2 language pairs and seems to be a universal feature that occurs throughout the course of foreign language learning (Ding et al., 2012; Graham & Post, 2018; Mennen & Leeuw, 2014; Mennen, 2015). In addition, the distinct pitch performance in L2 appears to suggest a differing degree of difficulty in acquiring Spanish intonation depending on sentence type, stress position and gender. Beyond this, our study extends previous hypotheses by proposing a progression in difficulty levels from the phonological to the phonetic dimension, suggesting that this difficulty ranking exists not merely when acquiring the L2 phonological contours, but also when implementing the pitch values of different sentence types.

From a teaching perspective, our study sheds light on L2 prosody learning, particularly on the L2 acquisition of Spanish intonation by Chinese learners which, despite its crucial importance, has not been addressed in many publications. Overall, it is proposed that there should be special training methods based on specific tasks to help reduce learners' foreign accents, as they showed distinct performances for L2 intonation patterns, according to the similarities and dissimilarities between the first and target language. Furthermore, our results suggest that the training program should not only include the phonological knowledge of target intonation contours but also, and perhaps more importantly, should allow learners to interact with language-appropriate contexts and to produce pitch implementation details in a native-like way. At this point, a growing number of recent investigations are devoted to the development of intonation teaching techniques. For instance, pitch gestures have been reported to benefit L2 prosody learning (Baills, 2016; Bernardis & Gentilucci, 2006; Gullberg, 2006; Jia & Wang, 2013a, 2013b; Morett & Chang, 2015), particularly in acquiring the low nuclear accents, which constitute the most difficult patterns for Chinese L2 learners. Other approaches, such as music training activities and speech visualization

tools, can also help learners to progress in L2 speech production. Nevertheless, considering that Spanish intonation has different levels of difficulty depending on sentence type, any pedagogical proposal should include scaffolding techniques in order to ease the way into the phonetic implementation.

Finally, some limitations should be noted in the current study. One such limitation is the lack of more precise criteria on the sample profile. Specifically, due to the dramatic reduction of possible samples, we did not specifically control for Chinese subjects' origin, age of L2 acquisition, or length of exposure to the target-language environment. Moreover, the sample sizes of male and female speakers were unequal across the three language groups. However, the effect of gender imbalance on the pitch range values was minimized by introducing the PDQ measure, which can effectively normalize the F0 variation data and make the LTD results more robust. Another limitation is related to the elicitation of the corpus. Due to the ambiguity of the situational contexts designed for the information-seeking *wh*-question, there were some unexpected nuclear pitch patterns in the L1 production, for instance, L+¡H* L%, which has been described as a form of the reintroduction of a previous topic rather than a request for new information. Finally, although our study found that there were some general pitch range deviations for L2 speakers, we did not discover exactly how the F0 range was realized depending on the syntactic position of the phrase and in which positions the learners deviated most from the L1 pitch patterns. Therefore, further research could take into account the effect of syntactic position to examine whether the increased pitch level and the compressed pitch span occur along with the overall L2 utterances or whether they are position-sensitive (e.g. increasing only on the low pitch targets while high tones remain basically unchanged).

8. Acknowledgments

This research was supported by a Ph.D. scholarship provided by China Scholarship Council (No.201808390059) to the first author, and by the postdoctoral program Juan de la Cierva awarded to the second author. Special thanks to Paolo Roseano who kindly revised the Sp_ToBI notation of the corpus. Also, we want to thank all Chinese and Spanish participants that allowed us to record their voices for this study.

9. References

Abu-Al-Makarem, A. & Petrosino, L. (2007) Reading and spontaneous speaking fundamental frequency of young Arabic men for Arabic and English languages: A comparative study. *Perceptual and Motor Skills*, 105(2), 572–580.

Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Magdalena, O.P. & Trouvain, J. (2014) Comparison of pitch range and pitch variation in Slavic and Germanic languages. In N. Campbell, D. Gibbon & D. Hirst (eds) *Proceedings of the 7th International Conference on Speech Prosody*. (pp. 776–780). Dublin, Ireland: Trinity College Dublin.

Atoye, R. O. (2005) Non-native perception and interpretation of English intonation. *Nordic Journal of African Studies*, 14(1), 26-42.

Backman, N. (1979) Intonation Errors in Second-Language Pronunciation of Eight Spanish-Speaking Adults Learning English. *Interlanguage Studies Bulletin*, 4(2), 239–265.

Baills, F. (2016). *Producing pitch gestures enhances the acquisition of Mandarin Chinese tones*. Unpublished MA thesis. Departament of Translation and Language Science, Universitat de Ponpeu Fabra.

Barto, K. A. (2015). *Mandarin Speakers' Intonation in their L2 English*. Unpublished doctoral dissertation. Program in Second Language Acquisition and Teaching, University of Arizona.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Beckman, M. E. & Venditti, J. J. (2011) Intonation. In J. Goldsmith, J. Riggle & A. C. L. Yu (eds) *The handbook of phonological theory*. (pp. 485–532). Oxford, UK: Blackwell.

Benesty, J., Chen, J., Huang, Y. & Cohen, I. (2009) Pearson correlation coefficient. In I. Cohen, Y. Huang, J. Chen & J. Benesty (eds) *Noise reduction in speech processing*. (pp. 1–4) Heidelberg, Germany: Springer.

Bernardis, P. & Gentilucci, M. (2006) Speech and gesture share the same communication system. *Neuropsychologia*, 44(2), 178–190.

Billmyer, K. & Varghese, M. (2000) Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Applied Linguistics*, 21(4), 517–552.

Bowen, J. D. (1956) A comparison of the intonation patterns of English and Spanish. *Hispania*, 39(1), 30–35.

Brebner, J. (2003) Gender and emotions. *Personality and Individual Differences*, 34(3), 387–394.

Busà, M. G. & Urbani, M. (2011) A Cross Linguistic Analysis of Pitch Range in English L1 and L2. In W.-S. Lee & E. Zee (eds) *Proceedings of the 17th International*

Congress of Phonetic Sciences. (pp. 380–383). Hong Kong, China: City University of Hong Kong.

Cadierno, T., Hansen, M., Lauridsen, J. T., Eskildsen, S. W., Fenyvesi, K., Jensen, S. H. & aus der Wieschen, M. V. (2020) Does younger mean better? Age of onset, learning rate and short-term L2 proficiency in young Danish learners of English. *Vigo International Journal of Applied Linguistics*, 17, 57–86.

Chao, Y. R. (1968) *A Grammar of Spoken Chinese*. Beijing: The Commercial Press.

Chen, G. T. (1972). *A comparative study of pitch range of native speakers of Midwestern English and Mandarin Chinese: An acoustic study*. Unpublished doctoral dissertation. Department of Philosophy, University of Wisconsin-Madison.

Chen, S. H. (2005) The effects of tones on speaking frequency and intensity ranges in Mandarin and Min dialects. *The Journal of the Acoustical Society of America*, 117(5), 3225–3230.

Chen, Y. & Gussenhoven, C. (2008) Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36(4), 724–746.

Cook, V. (2012) The native speaker and Multicompetence. In P. Robinson (ed) *Routledge Encyclopedia of Second Language Acquisition*. (pp. 447-450). London, UK: Routledge.

___. (2016) Premises of multicompetence. In V. Cook & W. Li (eds) *The Cambridge Handbook of Linguistic Multicompetence*. (pp. 1-25). Cambridge: Cambridge University Press.

Cortés Moreno, M (1997). Sobre la percepción y adquisición de la entonación española por parte de hablantes nativos de chino. *Estudios de Fonética Experimental*, 9, 67–134.

___. (1999) *Adquisición de la entonación española por parte de hablantes nativos de chino*. Unpublished doctoral dissertation. Departamento de Didáctica de la Lengua y la Literatura, Universitat de Barcelona.

___. (2001) Percepción y adquisición de la entonación española en enunciados de habla espontánea: el caso de los estudiantes taiwaneses. *Estudios de Fonética Experimental*, 11, 90–119.

___. (2004) Análisis acústico de la producción de la entonación española por parte de sinohablantes. *Estudios de Fonética Experimental*, 13, 80–110.

___. (2005) Análisis experimental del aprendizaje de la acentuación y la entonación españolas por parte de hablantes nativos de chino. *Phonica*, 1, 1–25.

Cruttenden, A. (1997) *Intonation*. Cambridge: Cambridge University Press.

Cuenca, M. J. (1997) Form-use mappings for tag questions. *Amsterdam Studies in the Theory and History of Linguistic Science*, 4, 3–20.

Ding, H., Hoffmann, R. & Hirst, D. (2016) Prosodic transfer: A comparison study of f0 patterns in L2 English by Chinese speakers. *Proceedings of the 8th International Conference on Speech Prosody* (pp. 756–760). Boston, USA: Boston University.

Ding, H., Jokisch, O. & Hoffmann, R. (2012) A phonetic investigation of intonational foreign accent in Mandarin Chinese learners of German. In Q. Ma, H. Ding & D. Hirst (eds) *Proceedings of the 6th International Conference on Speech Prosody.* (pp. 374–377). Shanghai, China: Tongji University Press.

Elvira-García, W., Roseano, P., Fernández-Planas, A. M. & Martínez-Celdrán, E. (2016) A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan. *Language Resources and Evaluation*, 50(4), 767–792.

Estebas-Vilaplana, E. & Prieto, P. (2010) Castilian Spanish intonation. In P. Prieto & P. Roseano (eds) *Transcription of intonation of the Spanish language*. (pp. 17–48). München, Germany: Lincom Europa.

Estebas-Villaplana, E. (2008) The evaluation of intonation: pitch range differences in English and in Spanish. In G. Thompson & L. Alba-Juez (eds) *Evaluation in Context*. (pp. 285–301) Amsterdam, The Netherlands: John Benjamins.

Face, T. L. & Prieto, P. (2007) Rising accents in Castilian Spanish: A revision of Sp_ToBI. *Journal of Portuguese Linguistics*, 6(1), 117–146.

Félix-Brasdefer, J. C. (2010) Data collection methods in speech act performance: DCTs, role plays, and verbal reports. *Speech Act Performance: Theoretical, Empirical and Methodological Issues*, 26, 41-56.

Font-Rotchés, D. (2005). *L'entonació del català. Patrons melòdics, tonemes i marges de dispersió.* Unpublished doctoral dissertation. Department of Didactics and Literature, Universitat de Barcelona.

Frota, S. & Prieto, P. (2015) *Intonation in Romance*. Oxford: Oxford University Press.

Girding, E., Zhang, J. & Svantesson, J.-O. (1983) A generative model for tone and intonation in Standard Chinese. *Working Papers (Lund University: Department of Linguistics)*, 25, 53-65.

Graham, C. & Post, B. (2018) Second language acquisition of intonation: Peak alignment in American English. *Journal of Phonetics*, 66, 1–14.

Gullberg, M. (2006) Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL-International Review of Applied Linguistics in Language Teaching*, 44(2), 103–124.

Gussenhoven, C. (2004) *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C. & Chen, A. (2000) Universal and language-specific effects in the perception of question intonation. In B. Yuan, T. Huang & X. Tang (eds) *Proceedings of the 6th International Conference on Spoken Language Processing*. (pp. 91–94). Beijing, China: China Military Friendship Publish.

Hanley, T. D., Snidecor, J. C. & Ringel, R. L. (1966) Some acoustic differences among languages. *Phonetica*, 14(2), 97–107.

Hewings, M. (1995) The English Intonation of Native Speakers and Indonesian Learners: a Coniparative Study. *RELC Journal*, 26(1), 27–46.

Hong, W. (2012) *An experimental study of Chinese students' English intonation pattern*. Unpublished doctoral dissertation. Department of Engllish Philosophy, Nankai University.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S. & Hothorn, M. T. (2016) Package 'multcomp': Simultaneous Inference in General Parametric Models. *Project for Statistical Computing*. Vienna, Austria.

Hualde, J. I. (2003) El modelo métrico y autosegmental. In P. Prieto (coord) *Teorías de La Entonación*. (pp. 155–184). Barcelona, Spain: Ariel.

Hualde, J. I. & Prieto, P. (2015) Intonational variation in Spanish: European and American varieties. In S. Frota & P. Prieto (eds) *Intonation in Romance*. (pp. 350–391). Oxford, UK: Oxford University Press.

Hyman, L. M. (2006) Word-prosodic typology. *Phonology*, 23(2), 225–257.

___. (2012) In defense of prosodic typology: A response to Beckman and Venditti. *Linguistic Typology*, 16(3), 341–385.

Ji, X. (2010) *Acquisition of intonation by Chinese EFL learners–An empirical study based on experimental phonetics*. Unpublished MA thesis. Department of English Linguistics, Zhejiang University.

Ji, X., Wang, X. & Li, A. (2009) Intonation patterns of yes-no questions for Chinese EFL learners. *Proceedings of the 2009 Oriental COCOSDA International Conference on Speech Database and Assessments*. (pp. 88–93). Urumqi, China: IEEE.

Ji, X., Xia, W., Yuan, J. & Li, A. (2012) Intonation patterns of *wh*-questions from Chinese EFL learners. In Q. Ma, H. Ding & D. Hirst (eds.) *Proceedings of the*

6th International Conference on Speech Prosody. (pp. 741–744). Shanghai, China: Tongji University Press.

Jia, L. & Wang, J. Q. (2013a). On the effects of visual processing on tone production by English-speaking learners of Chinese. *TCSOL Studies*, 52(4), 30–34.

___. (2013b.) The effects of visual processing on tone perception by native English-speaking learners of Chinese. *Chinese Teaching in the World*, 27, 548–557.

Jia, Y., Xiong, Z. & Li, A. (2006) The Effect of Focal Accents upon Sentential Pitch in Standard Chinese. *Report of Phonetic Research*, 61-68. Beijing, China.

Jilka, M. (2000) Testing the contribution of prosody to the perception of foreign accent. *New Sounds*, 4, 199–207.

Juan-Garau, M. & Pérez-Vidal, C. (2007) The effect of context and contact on oral performance in students who go on a stay abroad. *Vigo International Journal of Applied Linguistics*, 4, 117–134.

Juffs, A. (1990) Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *IRAL-International Review of Applied Linguistics in Language Teaching*, 28(2), 99–118.

Jun, S.-A. (ed) (2006) *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press.

Keating, P. & Kuo, G. (2012) Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, 132(2), 1050–1060.

Kharkhurin, A. V. (2008) The effect of linguistic proficiency, age of second language acquisition, and length of exposure to a new cultural environment on bilinguals' divergent thinking. *Bilingualism: Language and Cognition*, 11(2), 225–243.

Kochanski, G. & Shih, C. (2003) Prosody modeling with soft templates. *Speech Communication*, 39(3–4), 311–352.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017) lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1-27.

Ladd, D. R. (1996) *Intonational phonology*. Cambridge: Cambridge University Press.

___. (2001) Intonational universals and intonational typology. In M. Haspelmath, E. König, W. Oesterreicheran & W. Raible (eds) *Language Typology and Language Universals: An International Handbook*. (pp. 1-18). Berlin, Germany: De Gruyter.

Lenth, R., Singmann, H. & Love, J. (2019) Emmeans: Estimated marginal means, aka least-squares means. *R Package Version 1.3.4*.

Ling, F. (2003). *A Study on the acoustic correlates of emphasized Tone 3 in Mandarin Chinese*. Unpublished doctoral dissertation. Deparment of linguistics, Peking University.

Liu, F. & Xu, Y. (2005) Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, 62(2–4), 70–87.

Liu, Y. H. (2005) La entonación del español hablado por taiwaneses. *Biblioteca Phonica*, 2, 1–807.

Mackey, W.F. (2000) The description of bilingualism. In L. Wei (ed) *The Bilingualism Reader*. (pp.26-54). Oxford, New York: Routledge.

Majewski, W., Hollien, H. & Zalewski, J. (1972) Speaking fundamental frequency of Polish adult males. *Phonetica*, 25(2), 119–125.

May, S. (ed) (2013) *The multilingual turn: Implications for SLA, TESOL, and bilingual education*. New York: Routledge.

McGory, J. T. (1997). *Acquisition of intonational prominence in English by Seoul Korean and Mandarin Chinese speakers*. Unpublished doctoral dissertation. Deparment of Speech and Hearing Sciences,The Ohio State University.

Mennen, I., Schaeffler, F. & Docherty, G. (2007) Pitching it differently: a comparison of the pitch ranges of German and English speakers. In J. Trouvain & W. J. Barry (eds) *Proceedings of the 16th International Congress of Phonetic Sciences*. (pp. 1769–1772). Saarbrücken, Germany: Pirrot GmbH.

Mennen, I. & Leeuw, E. De. (2014) Beyond segments. *Studies in Second Language Acquisition*, 36(2), 183–194.

Mennen, I. (1998) Can language learners ever acquire the intonation of a second language? In *ESCA Workshop on Speech Technology in Language Learning (STiLL 98)*. (pp. 17–20). Marholmen, Sweden: Department of Speech, Music and Hearing, KTH.

___ (2015) Beyond segments: towards an L2 intonation learning theory (LILt). In E. Delais-Roussarie, M. Avanzi & S. Herment (eds) *Prosody and Language in Contact: L2 Acquisition, Attrition and Languages in Multilingual Situations*. (pp. 171–188). Berlin, Germany: Springer.

Mennen, I., Chen, A. & Karlsson, F. (2010) Characterising the internal structure of learner intonation and its development over time. In K. Dziubalska-Kołaczyk, M. Wrembel & M. Kul (eds) *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech*. (pp. 319–324). Poznań, Poland: Adam Mickiewicz University.

Mennen, I., Schaeffler, F. & Dickie, C. (2014) Second language acquisition of pitch range in German learners of English. *Studies in Second Language Acquisition*, 36(2), 303–329.

Mennen, I., Schaeffler, F. & Docherty, G. (2012) Cross-language differences in fundamental frequency range: A comparison of English and German. *The Journal of the Acoustical Society of America*, 131(3), 2249–2260.

Mennen, I. (2007) Phonological and phonetic influences in non-native intonation. In J. Trouvain & U. Gut (eds) *Non-Native Prosody: Phonetic Description and Teaching Practice*. (pp. 53–76). Berlin, Germany: De Gruyter Mouton.

Morett, L. M. & Chang, L.-Y. (2015) Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, 30(3), 347–353.

Ng, M. L., Hsueh, G., Leung, S. & Shing, C. (2010) Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology*, 12(3), 230–236.

Niebuhr, O. & Skarnitzl, R. (2019) Measuring a speaker's acoustic correlates of pitch–but which? A contrastive analysis based on perceived speaker charisma. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds) *Proceedings of the 19th International Congress of Phonetic Sciences*. (pp. 1774-1778). Melbourne, Australia: Australasian Speech Science and Technology Association Inc.

Nolan, F. (2003) Intonational equivalence: an experimental evaluation of pitch scales. In M. J. Solé, D. Recasens & J. Romero (eds) *Proceedings of the 15th International Congress of Phonetic Sciences*. (Vol. 771, pp. 2–5). Barcelona, Spain.

___. (2006) Intonation. In B. Aarts & A. McMahon (eds) *The Handbook of English Linguistics*. (pp. 433–457). Hoboken, USA: Blackwell Publishing Ltd.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41(1), 1–16.

___. (1994) The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols & J. J. Ohala (eds) *Sound symbolism*. (pp. 1337-1340). Cambridge, UK: Cambridge University Press.

Ortega, L. (2010) The Bilingual Turn in SLA. Plenary delivered at the *Annual Conference of the American Association for Applied Linguistics*. Atlanta, GA, March 6-9.

___. (2011) New trends in SLA research: Theories, methods, ethics. *Invited lecture at National Tsing Hua University*. Taiwan, June 8.

___. (2014) Ways forward for a bi/multilingual turn in SLA. In Stephen M. (ed) *The multilingual turn: Implications for SLA, TESOL and bilingual education*. (pp. 32-53). New York: Routledge.

Pandey, A. (2016). The multilingual turn: Implications for SLA, TESOL, and bilingual education. Stephen M. (ed). 2014. New York: Routledge. *World Englishes*, 35(3), 475-477.

Payà, M. & Vanrell, M. del M. (2005) Yes-no questions and echo questions intonation in Majorcan and Minorcan Catalan: A cross-dialectal comparison. In M. Zdrojkowski (ed) *Phonetics and Phonology in Iberia (PaPI)*. (pp. 20–21). Barcelona, Spain.

Peters, J. (2019) Fluency and speaking fundamental frequency in bilingual speakers of High and Low German. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds) *Proceedings of the 19th International Congress of Phonetic Sciences*. (pp. 1-5). Melbourne, Australia: Australasian Speech Science and Technology Association Inc..

Pfenninger, S. E. & Singleton, D. (2016). Age of onset, socio-affect and cross-linguistic influence: a long-term classroom study. *Vigo International Journal of Applied Linguistics*, 3, 147–179.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Unpublished doctoral dissertation. Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

Prieto, P. & Roseano, P. (2019) Current issues and challenges in Spanish intonational research. In S. Colina & M.-G. Fernando (eds) *The Routledge Handbook of Spanish Phonology*. (pp. 222–233). UK: Taylor & Francis.

Rendall, D., Vokey, J. R. & Nemeth, C. (2007) Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance*, 33(5), 1208–1219.

Roseano, P., González, M., Borràs-Comes, J. & Prieto, P. (2016) Communicating epistemic stance: How speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Processes*, 53(3), 135–174.

Schack, K. (2000) Comparison of intonation patterns in Mandarin and English for a particular Speaker. In K. M. Crosswhite & J. McDonough (eds) *Working Papers in the Language Science*. (pp. 24-55). Rochester, USA: University of Rochester.

S.J.Eady. (1982) Differences in the f0 patterns of speech: Tone language versus stress language. *Language and Speech*, 25, 29–42.

Shao, P., Jia, Y. & Li, A. (2011) Intonation patterns of exclamations of Chinese EFL learners from Zhejiang. *Proceedings of the 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. (pp. 74–78). Hsinchu City, Taiwan: IEEE.

Shi, S., Zhang, J. & Xie, Y. (2014) Cross-language comparison of F0 range in speakers of native Chinese, native Japanese and Chinese L2 of Japanese: Preliminary results of a corpus-based analysis. In M. Dong, J. Tao, H. Li, T. F. Zheng & Y. Lu (eds) *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing*. (pp. 241–244). Singapore: IEEE.

Shih, C. (1988) Tone and intonation in Mandarin. *Working Papers, Cornell Phonetics Laboratory*, 3, 83–109.

J't Hart, J., Collier, R. & Cohen, A. (1990) *A Perceptual Study of Intonation: An Experimental Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Talkin, D. (1995) A robust algorithm for pitch tracking (RAPT). In W. B. Klejin & K. K. Paliwal (eds) *Speech coding and synthesis*. (pp. 495–518). Amsterdam, The Netherlands: Elsevier Science B.V.

Tanaka-Ishii, K. & Terada, H. (2011) Word familiarity and frequency. *Studia Linguistica*, 65(1), 96–116.

Team, R. C. (2016) R: *A language and environment for statistical computing* [Computer software manual]. Vienna, Austria.

Todaka, Y. (1990). *An error analysis of Japanese students' intonation and its pedagogical applications*. Unpublished MA thesis. Department of Lingsuitics, University of California.

Ullakonoja, R. (2007) Comparison of pitch range in Finnish (L1) and Russian (L2). In J. Trouvain & W. J. Barry (eds) *Proceedings of the 16th International Congress of Phonetic Sciences*. (pp. 1701–1704). Saarbrücken, Germany: Pirrot GmbH.

Urbani, M. (2012) Pitch range in L1/L2 English. An analysis of f0 using LTD and linguistic measures. In M. G. Busà & S. Antonio (eds) *Methodological perspectives on Second Language prosody*. (pp. 79–83). Padova, Italia: Università di Padova

Valdés, G., Poza, L. & Brooks, M. D. (2015) Language acquisition in bilingual education. In Wayne E. W., Sovicheth B. & Ofelia G. (eds) *The handbook of bilingual and multilingual education*. (pp. 56-74). Malden, MA: Wiley-Blackwell.

Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, 38(3), 253–265.

Van Dommelen, W. A. & Moxness, B. H. (1995) Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech*, 38(3), 267–287.

Vanrell, M. del M., Mascaró, I., Torres-Tamarit, F. & Prieto, P. (2010) When intonation plays the main character : information-vs.confirmation-seeking questions in Majorcan Catalan. *Proceedings of the 15th International Conference on Speech Prosody*. (pp. 4–7). Chicago, USA: S. Karger AG.

Vilaplana, E. E. (2008) Modelling final declarative intonation in English and Spanish. In T. Gibert & L. Alba (eds) *Estudios de Filología Inglesa: Homenaje a La Dra. Asunción Alba Pelayo*. (pp. 109–129). Madrid, Spain: Ediciones UNED.

Volín, J., Poesová, K. & Weingartová, L. (2015) Speech melody properties in English, Czech and Czech English: Reference and interference. *Research in Language*, *13*(1), 107–123.

Wennerstrom, A. (1998) Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition*, 20(1), 1–25.

Wu, Z. (1982) Rules of intonation in Standard Chinese. *Preperation for Working Group on Intonation, 8th International Congress of Linguists*. (pp. 95–108). Tokyo, Japan.

Xu, H. (2009). *A survey study of Chinese EFL learners' acquisition of English intonation: A functional perspective*. Unpublished MA thesis. Department of English Philosophy, Jiangsu University.

___. (2005) Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46(3–4), 220–251.

___. (2017) Intonation in Chinese. In W. Behr, Y. Gu, Z. Handel, C.-T. J. Huang & J. Myers (eds) *Encyclopedia of Chinese language and linguistics*. (pp. 458–466). Boston: Brill.

Xu, Y. & Wang, Q. E. (2001) Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33(4), 319–337.

Yamazawa, H. & Hollien, H. (1992) Speaking fundamental frequency patterns of Japanese women. *Phonetica*, 49(2), 128–140.

Yip, M. (2002) *Tone*. Cambridge: Cambridge University Press.

Yuan, C., González-Fuente, S., Baills, F. & Prieto, P. (2019) Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Studies in Second Language Acquisition*, 41(1), 5–32.

Yuan, J. (2004). *Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling*. Unpublished doctoral dissertation. Department of lingusitics, Cornell University.

___. (2006) Mechanisms of question intonation in Mandarin. In B. Ma, Huo Q., E.-S. Cheng & H. LI (eds) *Proceedings of the 5th International Symposium on Chinese Spoken Languag*. (Vol. 4274, pp. 19–30). Berlin, Germany: Springer.

Yuan, J., Dong, Q., Wu, F., Luan, H., Yang, X., Lin, H. & Liu, Y. (2018) Pitch characteristics of L2 English speech by Chinese speakers: A large-scale study. *Proceedings of the 19th Annual Conference of the International Speech Communication Association*. (pp. 2593–2597). Hyderabad, India: Curran Associates, Inc.

Yuan, J. & Liberman, M. (2014) F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, 65, 67–74.

Yuan, J. & Shih, C. (2004) Confusability of Chinese Intonation. In B. Bernardl & M. Isabelle (eds) *Proceedings of the 2nd International Conference on Speech Prosody*. (pp. 1–4). Nara, Japan.

Zimmerer, F., Jügler, J., Andreeva, B., Möbius, B. & Trouvain, J. (2014) Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers. In N. Campbell, D. Gibbon & D. Hirst (eds) *Proceedings of the 7th International Conference on Speech Prosody*. (pp. 1037–1041). Dublin, Ireland: Trinity College Dublin.

Appendix A

Biographical information for L2 learners in the study, with Mandarin Chinese as the L1.

Code	Gender	Age at test	Age of acquisition	Month of residence in Spain	L2 proficiency	China origin	L2 dialect
ci01	F	22	18	10	B2	Henan	PS
ci02	F	22	18	10	B2	Sichuan	PS
ci03	M	26	21	12	B2 (SJ)	Shandong	PS
ci04	F	22	18	12	B2	Liaoning	PS
ci05	M	23	19	13	B2	Jiangsu	PS
ci06	F	23	19	14	B2	Jiangsu	PS
ci07	F	24	21	20	B2 (SJ)	Liaoning	PS
ci08	F	22	18	22	B2	Jiangsu	PS
ci09	M	31	24	22	B2	Shandong	PS
ci10	F	22	18	22	B2	Zhejiang	PS
ci11	F	23	17	25	B2	Anhui	PS
ci12	M	24	18	25	B2	Jiangxi	PS
ci13	M	24	18	25	B2	Shanxi	PS
ci14	M	21	19	32	B2 (SJ)	Guangdong	PS
ci15	F	23	19	40	B2 (SJ)	Ningxia	PS
ci16	F	23	18	60	B2 (SJ)	Shanghai	PS
ca17	F	24	18	1	C1	Shandong	PS

ca18	F	22	19	2	C1	Gansu	PS
ca19	F	24	20	2	C1	Jilin	PS
ca20	F	23	19	3	C1	Neimenggu	PS
ca21	F	22	18	7	C1	Beijing	PS
ca22	F	22	18	8	C1	Shanghai	PS
ca23	F	24	19	10	C1	Shanxi	PS
ca24	F	22	18	11	C1	Heilongjiang	PS
ca25	F	26	24	24	C1	Guangxi	PS
ca26	F	25	18	25	C1	Heilongjiang	PS
ca27	F	30	19	25	C2	Liaoning	PS
ca28	F	24	22	25	C1	Zhejiang	PS
ca29	F	29	18	36	C1	Henan	PS
ca30	F	24	18	40	C1	Heilongjiang	PS
ca31	F	29	19	40	C1	Anhui	PS
ca32	F	26	12	48	C2	Tianjin	PS

Note: Proficiency (SJ = self-judgement by L2 learners); L2 dialect (PS = Peninsular Spanish).

Appendix B

Test items

(1)	Sample contexts for information-seeking *yes-no* question: Entras en una frutería donde no has estado nunca y le preguntas al dependiente si tiene mandarinas. *¿Tiene mandarinas?*
(2)	Sample contexts for information-seeking *wh*-question: Has quedado con dos buenas amigas para hacer compras esta tarde. Pero al llegar la hora, sólo se presenta una chica y no ves a la otra amiga, Marina. Pregúntale dónde está Marina. *¿Dónde está Marina?*
(3)	Sample contexts for disjuntive question: Has invitado a buen amigo a tu piso para una cena. Después de acabar los platos principales, le preguntas si quiere tarta o helado de postre. *¿Quieres tarta o helado?*
(4)	Sample contexts for confirmation-seeking *yes-no* question: Estás hablando con tu compañero de piso y ves que está cubierto con dos mantas. Tú infieres que tu compañero de piso debe de tener frío y le preguntas si es así. *¿Tienes frío?*
(5)	Sample contexts for confirmation-seeking tag question: Vas a una discoteca con un amigo y ves que baila muy bien, así que supones que practica baile a menudo (mucho) y le preguntas si es así. *Bailas a menudo, ¿verdad?*

Other test items

Stress position	Penultimate stressed syllable	Final stressed syllable
	¿Tiene cambio? ¿Cuándo viene? ¿Viene el lunes o el martes? ¿Tienes hambre? Sales al mercado, ¿no?	¿Tiene melón? ¿Dónde está Manuel? ¿Quieres melón o melocotón? ¿Tienes calor? No te encuentras bien, ¿verdad?

Social attitudes, intelligibility and comprehensibility: The role of the listener in the perception of non-native speech

Ellen Simon
Department of Linguistics
Ghent University, Belgium
Ellen.Simon@UGent.be

Chloé Lybaert
Department of Translation, Interpreting and Communication
Ghent University, Belgium
Chloe.Lybaert@UGent.be

Koen Plevoets
Department of Translation, Interpreting and Communication
Ghent University, Belgium
Koen.Plevoets@UGent.be

Abstract

This study investigates the role of the listener in the perception of non-native speakers and their speech. Our goal is to examine the impact of listener characteristics on their attitudes towards non-native speakers and their speech. In addition, we aim to explore the relationship between listeners' attitudes and the intelligibility, comprehensibility and perceived foreign accentedness of non-native speech. 126 participants living in Flanders, Belgium, completed a questionnaire probing their attitudes towards non-native speakers of Dutch. An analysis of background variables of these listeners revealed that their age, educational level, extent of contact with non-native speakers and especially political preference could predict their responses to questionnaire items. In a subsequent session, participants performed (a) a speaker/speech evaluation task, (b) a transcription task measuring intelligibility, and (c) a comprehensibility and accentedness rating task. This latter session required participants to transcribe and evaluate speech samples of Dutch produced by speakers of Mandarin Chinese. We found a significant correlation between comprehensibility and accentedness and a number of attitudinal dimensions, such as the perceived status of, and solidarity with, the speaker. The study has implications for language testing, as it demonstrates the impact of listeners' social attitudes on the assessment of non-native speech.

Keywords: Intelligibility; comprehensibility; social attitudes; non-native speech: foreign accent.

Abstract

Diese Studie untersucht die Rolle des Zuhörers bei der Wahrnehmung von Nicht-Muttersprachlern und ihrer Rede. Wir wollten den Einfluss von Zuhörermerkmalen auf die Einstellung der Zuhörer gegenüber Nicht-Muttersprachlern und Nicht-Muttersprachlern untersuchen und die Beziehung zwischen der Einstellung der Zuhörer und der Verständlichkeit ("intelligibility" sowie "comprehensibility") und wahrgenommenen Fremdsprachenakzentuierung von Nicht-Muttersprachlern untersuchen. Insgesamt 126 Teilnehmer in Flandern, Belgien, füllten einen Fragebogen aus, der ihre Einstellung gegenüber Nicht-Muttersprachlern untersuchte. Eine Analyse der Hintergrundvariablen dieser Zuhörer ergab, dass ihr Alter, ihr Bildungsniveau, das Ausmaß des Kontakts mit Nicht-Muttersprachlern und insbesondere ihre politischen Präferenzen ihre Antworten auf die Fragen des Fragebogens vorhersagen konnten. In einer anschließenden Sitzung führten die Teilnehmer (a) ein Sprecher/Sprache-Evaluierungsexperiment, (b) eine Transkriptionsaufgabe zur Messung der Verständlichkeit und (c) eine Begreifbarkeits- und Akzentuierungsbewertung durch, bei der Sprechproben ausgewertet wurden, die von niederländischsprachigen Zweitsprachlern des Mandarin-Chinesischen erstellt worden waren. Wir fanden eine signifikante Korrelation zwischen Verständlichkeit und Akzentuierung und einer Reihe von Einstellungsdimensionen, wie z.B. dem wahrgenommenen Status des Sprechers und der Solidarität mit ihm. Die Studie hat potentielle Auswirkungen auf Sprachtests, da sie den Einfluss der sozialen Einstellungen der Zuhörer auf die Beurteilung von Nicht-Muttersprachlern.

Keywords: Verständlichkei; Begreifbarkeit; soziale Einstellunge; Nicht-Muttersprachler; ausländischer Akzent.

1. Introduction

It is commonly known that non-native speech typically contains acoustic-phonetic properties that deviate from average native speech properties. In fact, late learners, broadly defined as those who learn a non-native language well after childhood, are rarely indistinguishable from those who have acquired the language from birth (exceptions are reported in Bongaerts, Mennen & van der Slik, 2000). This is the result of cross-linguistic interaction, the process through which speakers transfer phonetic and phonological properties, both segmental and prosodic, from their

native language into the target language and vice versa (Major, 2001). Cross-linguistic interaction is the central concept in second language speech and perception models, such as the Speech Learning Model (Flege, 1995) and the Perceptual Assimilation Model (Best, 1995; Best & Tyler, 2007). More specifically, a growing number of studies have shown that acquiring phonological properties of a non-native language which are not contrastive in the native language remains problematic, even for advanced learners, and this holds for both production and perception (see Bohn & Munro, 2007, for an overview). Awareness of these findings, as well as the gradual realization that the concept of the native speaker is problematic in many contexts (cf. the growing recognition of World Englishes and the related issue of ownership, Davies, 2013), have led to a shift of pronunciation goals in foreign language learning. The goal of a native-like pronunciation, based on the model of the native speaker as the ultimate target, has gradually been replaced by the goal of fluent communication and mutual intelligibility (Jenkins, 2000; 2005; Levis & LeVelle, 2010).

In line with this shift, oral proficiency targets in foreign language learning are now increasingly expressed in terms of intelligibility. Examples can be found in the descriptions of proficiency levels in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), where the extent to which learners are intelligible to interlocutors plays a key role in the descriptions of the proficiency scales (Beinhoff, 2014). Intelligibility is, however, a concept that cannot easily be defined, because it is the result of an intricate interplay between different factors related not only to the speaker but also—crucially—to the listener (see Section 2.1). There is, indeed, growing awareness of the role of the listener in studies on intelligibility. It is now generally acknowledged that the extent to which a speaker is intelligible to the interlocutor is ultimately not only determined by the speaker's pronunciation and lexicogrammar but also by characteristics of the listener. It is this role of the listener that is the focus of the present study. In particular, we examine the impact of listener characteristics on the formation of social attitudes towards non-native speakers and their speech. The aim is to find out whether listeners' characteristics can predict social attitudes towards non-native speakers and non-native speech and whether listeners' evaluations can predict the intelligibility, comprehensibility and (foreign) accentedness of non-native speech (see Section 3).

Section 2 contains a literature review. We first focus on the differences and similarities between three related constructs: "intelligibility", "comprehensibility" and "accentedness". We briefly explain how these constructs have been defined in earlier research and which definitions we use in the present paper (Section 2.1). We then look at the role of the listener in earlier research on intelligibility and comprehensibility of non-native speech (Section 2.2). The specific research questions in our own empirical

study are discussed in Section 3. We describe the methodology in Section 4 and the results in Section 5. Section 6 contains the discussion. Conclusions and suggestions for further research are presented in Section 7.

2. Review of the Literature

2.1. Intelligibility, comprehensibility and accentedness: terminology and measurement

In the present study, we follow Munro & Derwing (1995a) in their identification of three separate but related constructs: "intelligibility", "comprehensibility" and "accentedness". Munro & Derwing (1995a) distinguish between "intelligibility", which they define as the extent to which actual speech units are recognized by the listener, and "(perceived) comprehensibility", which refers to the extent to which speech is reported to be understood by the listener. The two main linguistic dimensions that have been shown to influence a speaker's intelligibility and comprehensibility are pronunciation (segmental properties, prosody and fluency) and lexicogrammar (richness of vocabulary as well as accuracy and complexity of grammar) (Saito, Trofimovich & Isaacs, 2017). Although intelligibility and comprehensibility are related in the sense that speech which is highly intelligible is also generally rated high for comprehensibility, Munro & Derwing (1995a, b) and Derwing & Munro (1997) showed that the two constructs are also partially independent of each other. Surprisingly, it appeared to be the case that a number of speakers who were almost perfectly intelligible to some listeners (as shown by their near perfect transcription scores) were not rated as highly comprehensible by these same listeners. Intelligibility and comprehensibility are often discussed in relation to a third concept, that of (foreign or non-native) "accentedness". Gallardo del Puerto, García Lecumberri & Gómez Lacabex (2015) define a foreign accent as "the pronunciation of a language which presents phonological features that deviate from native phonological usage" (p. 203). Accentedness, as a characteristic of non-native speech, is closely related to intelligibility, as it presents a deviation from listeners' expectations and hence a potentially adverse condition for optimal speech perception (Mattys et al., 2012). However, it is also closely linked to comprehensibility, in that the extent to which speech is judged to be ("foreign") accented is a matter of the listener's reported perception.

Since intelligibility refers to the actual recognition of specific units of speech (Kachru & Smith, 2008), whereas comprehensibility and accentedness refer to listeners' reported perceptions, the way in which these constructs are measured is different. Measures of intelligibility focus on the extent to which listeners are able to

correctly identify words or speech sounds targeted by the speaker. Kent & Miolo (1994) review several procedures that have been used in previous studies to measure speech intelligibility in children who are likely to be diagnosed with a speech disorder. These procedures include, for instance, attesting the percentage of correctly recognized target words, or more specifically consonants or vowels (embedded in consonantal frames, as in e.g. *feel-fill-fall-fell*). Such methods are similar to those used in studies on multilingual speakers or second language learners. Kang, Thomson & Moran (2018) compare five intelligibility measures in different varieties of non-native speech, in order to determine the efficacy of these measures in predicting listening comprehension scores. Measures included responses to true/false statements, scalar ratings of speech, perception of nonsense sentences, perception of filtered sentences, and transcription of speech. The measures differed in the amount and type of context that the listener received. Results revealed that the relationship between the five measures was weak, suggesting that they tap into slightly different constructs. The measure that best predicted listener comprehension of longer speech samples of the same speakers was a transcription task. This task required listeners to transcribe missing words in grammatically correct, but semantically nonsensical sentences, such as "The tall kiss can draw with an oak". The advantage of using nonsensical sentences over semantically meaningful sentences lies in the absence of context which would enable listeners to guess the missing words. As Kang et al. (2018: 119) point out, the technique was developed for speech technology and audiology and was later adapted for use in studies on speech intelligibility (e.g. Bent & Bradlow, 2003).

As pointed out above, comprehensibility and foreign accentedness are different from intelligibility, in that they reflect listeners' reports on, respectively, the extent to which they understand the speaker and the extent to which the speech deviates from what they view as a native language accent. Comprehensibility and foreign accentedness are therefore typically measured through rating scales, in which listeners are asked to indicate the degree of comprehensibility and the strength of the foreign accent (cf. e.g. Munro & Derwing, 1995a; Munro, 2008).

2.2. The role of listener characteristics in intelligibility research

As noted in the Introduction, the role of the listener is increasingly recognised as crucial in studies on intelligibility, comprehensibility and accentedness. It is now generally acknowledged that the extent to which a speaker will be intelligible to the interlocutor will ultimately be determined not only by the speaker's pronunciation and lexicogrammar but also by characteristics of the listener. Rajaduraj (2007) refers to Morley (1991), who claims that "intelligibility may be as much in the mind of the listener as in the mouth of the speaker" (p. 499). A recent and comprehensive

study by Saito et al. (2019) examines how comprehensibility and foreign accentedness judgements are affected by listener characteristics, including the listener's first language (L1), level of proficiency in the second language (L2), age, experience, familiarity and metacognition. They compared responses to listener background questionnaires by 120 respondents (110 L2 listeners and 10 L1 listeners) with comprehensibility ratings of 50 speakers. The results showed that metacognition (level of awareness of the importance of "comprehensible" pronunciation and lexicogrammar), experience (extent of contact with L2 speakers) and L1-L2 distance predicted listeners' comprehensibility scores assigned to the L2 speakers. The effect of the listener's native language on L2 speech intelligibility was also studied, and Bent & Bradlow (2003) observed an "interlanguage speech intelligibility benefit" (ISIB). This means that non-native listeners were as accurate in recognizing words produced by a high-proficiency non-native speaker with whom they shared the native language as they were in recognizing words produced by a native speaker (cf. also Hayes-Harb, Smith, Bent & Bradlow, 2008). Winke, Gass & Myford (2013) examined the effect of accent familiarity, defined as the result of having studied the speaker's L1. For example, native English speakers who had studied Korean were considered to be familiar with a Korean accent in English. By examining comprehensibility and speech ratings of 107 listeners judging speech samples produced by 72 test takers, they found that the linguistic background of listeners should indeed be taken into account as a potential rater effect.

Besides listener characteristics such as his or her first language, accent familiarity and the extent of contact with L2 speakers, another aspect of listeners' profiles that has been put forward in intelligibility and comprehensibility research relates to listeners' social attitudes (see Kennedy & Trofimovich, 2019, for a recent discussion on social influences on comprehensibility). We define social attitudes as listeners' attitudes towards social stimuli, in this case non-native speakers and non-native speech. Rajadurai claims that while positive attitudes towards speakers and accents may enhance intelligibility and facilitate communication, listeners' biases or negative judgements may act as barriers to intelligibility (2007: 90). This claim is supported by a set of previous studies, some dating back to the '80s and '90s, but others just recently published (Babel & Russell, 2015; Eisenstein & Verdi, 1985; Lindeman, 2002; Rubin, 1992; Sheppard, Elliott & Baese-Berk, 2017; Taylor Reid, Trofimovich & O'Brien, 2019). These studies have shown that listeners' attitudes towards, and evaluations of, certain accents can be associated with intelligibility and/or comprehensibility. The studies widely differ in the methodological choices they make with respect to the type of task and stimuli and the precise nature of the construct they measure, ranging from intelligibility (e.g. Babel & Russell, 2015) over comprehensibility (e.g. Taylor Reid et al., 2019) to, more generally, perceived success of the interaction (e.g. Lindeman, 2002).

Babel & Russell (2015), for instance, showed that speech intelligibility decreased when listeners were shown a picture of an ethnically Chinese Canadian speaker compared to when no visual prime was presented. This effect was not found for White Canadian speakers, suggesting that listeners associate Chinese faces with less intelligible, non-native accents.

Taylor Reid, Trofimovich & O'Brien (2019) manipulated listeners' social biases before they were asked to evaluate second language speech. Listeners who were positively oriented towards Quebec French speakers of L2 English by hearing positive comments about them rated their comprehensibility more favourably than baseline listeners, who had not been presented with any comments. Negatively biased listeners showed divergent patterns of ratings: while negatively oriented younger listeners rated the speakers' comprehensibility more favourably, older negatively oriented listeners downgraded the speakers compared to the baseline group. The authors surmise that positively oriented listeners and younger negatively oriented listeners may have experienced enhanced solidarity with the L2 speakers (e.g. by drawing parallels to themselves when speaking an L2), while older negatively oriented listeners may have perceived a large social distance between themselves and the speakers (Taylor Reid et al., 2019: 434).

Finally, Lindeman (2002) examined whether native English speakers who were assessed as having either positive or negative attitudes towards Koreans performed differently on an interactional map task with native Korean speakers of English as L2. She found that some of the speakers with negative attitudes used strategies problematizing their interlocutor's utterances or avoidance strategies (e.g. not asking additional questions). Lindeman also found a direct relationship between attitudes and perceived success of the interaction, with more positive attitudes correlating with the perception of a higher success level.

In sum, although previous studies of the effect of listeners' social attitudes on how non-native speakers and their speech are perceived, have adopted different approaches and methods, there is a growing body of research suggesting that listeners' attitudes indeed impact on L2 speech perception and assessment and that they should therefore be taken into account in research on L2 intelligibility and comprehensibility. The question then arises as to which factors influence these attitudes. The literature on this topic mostly deals with large sets of respondents' background characteristics shaping their attitudes towards immigrants or refugees in general (e.g. Chandler & Tsai, 2001; Butkus, Maciulyte-Sniukiene & Matuzeviciute, 2016; Murray & Marx, 2013). The studies do not focus specifically on attitudes to non-native speakers and their speech. The respondents' background characteristics "age", "level of education", "extent of contact with the target population" and "political preference" are among the variables most often studied.

Butkus et al. (2016) review the literature on factors shaping attitudes towards immigrants and refugees and conclude that "in general, most positive and tolerant attitudes are associated with youth, high socio-economic status, high educational attainment and left-wing political sympathies" (Hainmueller & Hiscox, 2007). These factors are originally discussed in sociological theories, such as *Human Capital Theory*, which analyses the effect of education on attitudes. The general claim is that people with a higher level of education are more tolerant towards immigrants and refugees, because there is no economic competition with that group, whose members are often unskilled or semi-skilled (Mayda, 2006, in Butkus et al., 2016). Moreover, as pointed out by Chandler and Tsai (2001), people who have enjoyed a higher level of education tend to have more diverse and cosmopolitan networks, fostering tolerance. According to the *Contact Theory,* people who have closer contact with the target population have more positive attitudes to immigrants, since people are generally prejudiced towards things they do not know (Butkus et al., 2016: 287). Political preference is generally thought to be linked to attitudes towards migration in the sense that people with right-wing or conservative political ideologies tend to take a more negative stance towards immigrants (Chandler & Tsai, 2001; but see also Murray & Marx, 2013, who did not find an effect of political orientation). Finally, age is one of the demographic variables of respondents that is generally believed to have an effect on people's attitudes towards migration. Card, Dustman & Preston (2005) argue that there are several reasons for this. One explanation is that age marks the position of the individual in their economic cycle. For instance, people who are active on the labour market may show attitudes towards immigrants that are different from those of adolescents who do not yet need to compete for jobs. A large-scale survey of attitudes towards immigration in Europe showed that, if educational level is held constant, older people generally had stronger anti-immigrant views (Card, Dustman & Preston, 2005). The authors note that this difference between younger and older respondents may be the result of aging or of a difference between birth cohorts. In other words, it is possible that, as an individual gets older, they develop more negative attitudes towards immigrants, but it is also possible that people who grew up in, for instance, the 1940s developed less liberal attitudes towards immigration than people who grew up in the 1980s and kept these attitudes throughout their lives (Card, Dustman & Preston, 2005: 26). Further research is needed to tease out these factors.

3. The study

3.1. Aims and research questions

The goal of the present study was to examine the impact of listener characteristics on listeners' attitudes towards non-native speakers and non-native speech and to explore

the relationship between listeners' attitudes and the intelligibility, comprehensibility and perceived foreign accentedness of non-native speech. To this end, we set up experimental research in which native Dutch speakers first completed a background and attitudinal questionnaire. In a second session they were given a transcription task, a speaker/speech evaluation task and were asked to rate the comprehensibility and accentedness of L2 Dutch speech samples by native speakers of Mandarin Chinese. We opted for native speakers of Chinese because we wanted to use speech produced by speakers with a native language that is not related to Dutch (Dutch belongs to the Indo-European language family, while Chinese is part of the Sino-Tibetan language family) and hence possibly less intelligible than speech produced by speakers of a language closely related to Dutch (cf. also Caspers & Horloza, 2012).

We address the following two research questions:

1. Can listeners' characteristics (age, level of education, extent of contact with L2 speakers, political ideology) predict social attitudes towards non-native speakers and non-native speech?
2. Can listeners' evaluations of non-native speakers and non-native speech predict the intelligibility, comprehensibility and (foreign) accentedness of non-native speakers and non-native speech?

On the basis of the literature reviewed in Section 2.2., our hypothesis is that all four listener characteristics examined—age, level of education, extent of contact with L2 speakers and political ideology—will, to some extent, predict listeners' attitudes towards non-native speakers and non-native accents, as measured through listeners' responses in a questionnaire. Specifically, we predict that more lenient attitudes will be associated with younger respondents, respondents with a higher educational level, more contact with non-native speakers and with a preference for a more left-wing political ideology. By more lenient attitudes towards non-native speakers and their speech we mean, for instance, more patience with newcomers who have to learn Dutch, more positive expectations about conversations with non-native speakers and more tolerant attitudes towards non-native accented speech.

With respect to listeners' attitudes, as reflected in a non-native speaker evaluation experiment, our expectations are less pronounced, as previous studies addressing this issue have involved a variety of tasks and procedures. While we expect to find an association between listeners' evaluations of non-native speakers and speech on the one hand, and the intelligibility, comprehensibility and (foreign) accentedness scores of these listeners on the other, the effect is possibly a subtle one.

The study was carried out in Belgium, where Dutch, French and German are the three official languages, spoken in different regions of the country, with Brussels being officially bilingual in French and Dutch. The research was conducted in Flanders, the northern part of Belgium, where Dutch is the native language of the majority of the inhabitants, but where a multitude of other languages are used as home languages, as a result of migration. Most migrants have EU-nationalities, with migrants from Morocco, Syria and India as exceptions in the top 10 nationalities of migrants in 2016. In the list of non-EU-nationalities of international migrants in 2016, Chinese takes fifth position (see Section 3.2.2.2) (Noppe et al., 2018). Semyonov, Raijman & Gorofzeisky (2006) report fairly average levels of ethnic diversity and anti-immigrant sentiments in Belgium compared to 11 other Western European countries in the period 1988-2000. A 2017 survey by the Flemish government probing the beliefs and attitudes of the Dutch-speaking population in Flanders and Brussels revealed that more than half of the sample population believed that the presence of other cultures enriches society. By contrast, 14% of the population reported not to trust migrants and 30% believed that migrants are a threat to the Flemish culture and values (Noppe et al., 2018). Although our study focuses on non-native speakers instead of migrants, we predict that the same positive and negative attitudes reflected in these survey results will be echoed in our study.

3.2. Methodology

Our study consisted of two main parts, for which data were collected at two different times. An interval of two to three months between the two data collection points was opted for, because we wanted to prevent participants from being influenced by their answers in the first part (attitudes towards non-native speakers and their speech in general) when completing the second part (including a speaker evaluation task). We therefore describe the methodology of the two parts in separate sections (Section 3.2.1 and Section 3.2.2).

3.2.1. Part 1

For part 1, we designed a questionnaire with statements about non-native speakers and non-native speech.

3.2.1.1 Materials

The questionnaire was meant to probe participants' attitudes towards non-native speakers and non-native speech in general, i.e. not restricted to non-native speakers

or speech from a specific linguistic background. In order to do so, we designed 34 statements, covering nine themes: (1) attitudes towards learning Dutch, (2) political ideology, (3) view on second language acquisition processes, (4) willingness to put in extra effort when communicating with L2 speakers, (5) comprehensibility of L2 speakers, attitudes towards L2 speakers and their speech related to (6) status, (7) solidarity, and (8) aesthetic value (i.e., how beautiful a language is considered to be or how pleasant it is thought to sound), and (9) attitudes towards variation in language.

The statements were designed by the authors, but inspired by previous studies probing social attitudes towards language varieties and their speakers (Latour et al., 2012; Grondelaers & Speelman, 2013; Delarue, 2016; Lybaert, 2017). Nine themes were selected for which, on the basis of the literature, we had reason to believe they would be related to intelligibility and comprehensibility/accentedness of non-native speakers. Attitudes towards L2 speakers regarding status, solidarity and aesthetic value, for instance, are dimensions which are typically addressed in speaker evaluation experiments because they have been shown to be relevant factors (see e.g. Latour et al., 2012; Grondelaers & Speelman, 2013). "Attitudes towards variation in language" was included as a theme because attitudes towards non-standard variation have been the subject of much discussion in the Dutch-speaking area (e.g. Lybaert, 2017; Delarue, 2016) and we hypothesized that there may be a link between people's attitudes towards endogenous and exogenous varieties of Dutch. The theme "political ideology" was included because previous studies (reviewed in Section 2.2) claim that political sympathies are associated with attitudes towards immigrants and these, in turn, may affect intelligibility or comprehensibility of non-native speech. "Attitudes towards learning Dutch", "view on second language acquisition processes", "willingness to put in extra effort when communicating with L2 speakers" and "comprehensibility of L2 speaker" were not directly based on previous studies, but fitted our goal of probing participants' general beliefs about language acquisition and their stance towards conversations with non-native speakers. All statements are included in Section 4.1, Table 2.

Participants were asked to indicate on five-point Likert scales to what extent they agreed or disagreed with each statement (1=strongly disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, 5=strongly agree). None of the statements contained negations. Half of the statements (N=17) were formulated in a way that agreeing would reflect a fairly positive attitude (e.g. "I feel safe with non-native speakers of Dutch"), while agreement with the remaining half would be a sign of a rather negative attitude (e.g. "Speaking with a non-native speaker of Dutch often leads to misunderstandings").

At the end of the questionnaire, we also collected some background information: we included questions about the participants' age, gender, the place where they had

spent most of their childhood, and the highest qualification obtained. Participants were also asked about their political preference through a multiple-choice question with the options "far left", "left-of-centre", "centre", "right-of-centre" and "far right". To conclude, they were asked how much contact they have with non-native speakers of Dutch (rarely or never, on a monthly, weekly or daily basis). In this paper, we focus on the impact of the background variables "age", "highest qualification obtained", "political preference" and "contact with non-native speakers" on the attitudes of the respondents.

3.2.1.2. Procedure

The questionnaire was presented and completed through an online platform. Participants were instructed to complete the questionnaire on a PC or laptop, since we believed that completing a longer questionnaire on a smartphone might lead to concentration loss. The questionnaires were processed anonymously, and participants were invited to answer as honestly as possible. All participants were informed on their right to withdraw their answers at any point and on the anonymization of the data. They all completed an informed consent form.

3.2.1.3. Participants

A total of 126 respondents completed the questionnaire in Part 1. They were recruited from the researchers' social networks through social media. Only participants aged between 18 and 50 were encouraged to participate.

Responses to background questions in the questionnaire (cf. 3.1.2) revealed the following participant characteristics: the respondents' ages ranged between 18 and 50, with a mean age of 31. Of those respondents, 49 were male and 77 were female. They had spent most of their childhood in the regions East- and West-Flanders and Flemish Brabant (East-Flanders=89, West-Flanders=17, Flemish Brabant=17). Only three participants came from the provinces of Antwerp (N=2) and Limburg (N=1).

The highest qualifications obtained were a diploma in secondary education (vocational schooling (N=3), technical schooling (N=18) or general schooling (N=31)), a Bachelor's degree (N=56), a Master's or PhD degree (N=20). It should be noted that 37 participants were still studying and hence working towards a Bachelor's or Master's degree.

We observed the following distribution in the participants' political preference: far left (N=1), left-of-centre (N=21), centre (N=59), right-of-centre (N=44) and far

right (N=1). As only one respondent answered with "far right" and only one with "far left", these answers were added to the categories "right-of-centre" and "left-of-centre", respectively. When asked how much contact they had with non-native speakers of Dutch, the participants responded "rarely or never" (N=11), "monthly" (N=36), "weekly" (N=50) and "daily" (N=29).

3.2.2. Part 2

For the second part of the study, we used recorded speech samples from two non-native speakers. This second part of the study consisted of three tasks: a transcription task, a comprehensibility and accentedness judgement task, and a speaker/speech evaluation task. All tasks were offered to participants in LimeSurvey Version 2.73.1.

3.2.2.1. Participants

All participants of Part 1 (N=126) were also invited to participate in Part 2 of the study, two to three months after the data for Part 1 were collected. In total, 102 participants took part in the second part of the study. The remaining 24 participants did not react to the call for the second part of the study.

3.2.2.2. Speech stimuli

The second part of the study contained recorded speech samples from two non-native speakers of Dutch, henceforth referred to as speakers A and B. The speakers were selected for their very similar profiles: they were both female native speakers of Mandarin Chinese, who were born and raised in China, but moved to Belgium (in the area of Ghent) as adults. Both speakers have a Master's degree in Marketing (obtained in Belgium) and a native Dutch-speaking partner, with whom they (mainly) communicate in English. At the time of the recording, they had both reached CEFR proficiency level C1 in Dutch at the same centre for adult education in Ghent and were selected by their Dutch teacher as motivated learners. When asked whether interlocutors have difficulties understanding them when they speak Dutch, they both selected the answer "sometimes" (from the options "never", "rarely", "sometimes", "often" and "very often"). We opted for participants with a high proficiency level in Dutch, because we wanted to obtain information on the perception of L2 speakers who could perfectly express themselves in Dutch, but had a noticeable accent. Speech samples from two speakers were selected because we wanted to examine potential individual differences between two speakers with similar language profiles. We held the view that including more speakers for all three tasks would make the experiment

unnecessarily long, potentially leading to concentration loss on the part of the listeners.

Speaker A was 30 years old at the time of the recording and moved to Belgium seven years ago (age 23), when she started taking Dutch classes. She describes her Dutch as "quite good". Speaker B was 37 years old at the time of the recording. She moved to Belgium 13 years ago (age 24) but spent a total of four years in New York and Shanghai between her arrival in Belgium and the time of the recording. She perceives here own proficiency in Dutch as "average". The recordings were made in a sound-attenuated room. They consisted of three tasks: for task 1, informants read aloud 48 sentences; for task 2 they were asked to tell us something about the holiday of their dreams and for task 3 they read aloud a small text. They were given a few minutes to prepare this last task. The total time of the recording session was 45-60 minutes per person.

3.2.2.3. Tasks and materials

3.2.2.3.1. Transcription task

All participants completed a transcription task. This task contained a selection of 24 unique and grammatically correct Dutch sentences, 12 of which were produced by speaker A and 12 by speaker B. Each speaker had produced 6 semantically meaningful sentences and 6 nonsensical sentences. The audio sentences were presented together with their orthographic transcriptions, in which one target word was left out from every sentence. Participants were instructed to listen to the sentences and transcribe the missing words. Nonsensical sentences were included in order to avoid participants guessing from the semantic context which word was missing (cf. Kang et al., 2018; see Section 2.1). The missing target words contained six vowels, forming three vowel contrasts: /i/-/ɪ/, /e/-/ɛ/ and /a/-/ɑ/. Half of the missing words were nouns, the other half were verbs (see Appendix 1 for an overview of the selected sentences).

Table 1. Missing target words transcription task

		i	ɪ	e	ɛ	a	ɑ
Speaker 1	nonsensical	*piekt*	*pikt*	*heeft*	*heft*	*kaas*	*Kat*
	semantically meaningful	*biedt*	*bidt*	*peer*	*step*	*raad*	*Rat*
Speaker 2	nonsensical	*ziet*	*zit*	*zeep*	*schep*	*zaak*	*Zak*
	semantically meaningful	*stier*	*schip*	*veegt*	*vecht*	*kaapt*	*Kapt*

The respondents first had to transcribe the missing words in the semantically meaningful sentences: after two test sentences, the actual 12 semantically meaningful sentences of the two speakers were presented in random order. The respondents were given the instruction to write the word they heard under the sentence. If they really did not understand the word, they had to type "not understood". After filling in the missing words in the semantically meaningful sentences, the respondents were presented with the nonsensical sentences. Again, they were presented with two warm-up sentences first, followed by the actual 12 nonsensical sentences in random order.

3.2.2.3.2. Comprehensibility and accentedness judgement task

The second part of the test consisted of a comprehensibility and accentedness judgement task. In this task, participants were presented with an audio file of semi-spontaneous speech (of approximately 15 seconds; an orthographic transcription is included in Appendix 2) by each speaker (see Section 3.2.2.2). The order of the two speakers was randomized across participants. After each presentation of an audio file, they were asked to indicate on a five-point Likert scale to what extent they agreed with the statement "I think this person is easily intelligible" (*Ik vind deze persoon goed verstaanbaar*) and to select an answer to the question "How do you rate the non-native accent of this speaker?" (*Hoe sterk vind je het buitenlandse accent van deze persoon?*) from 1 ("no accent") to 5 ("very strong accent").

3.2.2.3.3 Speaker/speech evaluation task

The third task in Part 2 was a speaker/speech evaluation task. Participants were asked to listen to two identical excerpts of read speech (of approximately 20 seconds), a short narrative text, once produced by speaker A and once by speaker B (an orthographic transcription is included in Appendix 3). They were subsequently asked to indicate the extent of their (dis)agreement with 17 statements on the speech/speaker they had just heard. The 17 statements covered five themes: (1) aesthetic value of the speech, (2) status of the speaker, (3) solidarity with the speaker, (4) comprehensibility of the speaker and (5) perceived level of integration of the speaker in society. The statements were selected on the basis of existing speaker evaluation experiments (see e.g. Grondelaers & Speelman, 2013; Grondelaers & Van Gent, 2019). The order of both fragments was again randomized across participants (an overview of the statements is given in Section 4.1, Table 2).

3.2.3. Analysis

The 34 questions of Part 1 were first analysed with Principal Component Analysis in order to explore whether they could be grouped into broader categories, for example, according to the nine themes represented in the questionnaire. To this end, the 16 reversely coded questions were back-transformed , in order to align with the 18 other questions. However, many questions from different themes appeared to be correlated with each other. As a consequence, we decided not to work with principal components nor the nine underlying themes. Instead we analysed each of the 34 questions separately in a linear model with the six background variables as predictors (main effects). The results are presented in Section 4.1.

For the experiment in Part 2, the (102) respondents had to evaluate two speakers, so we fitted three Linear Mixed-effects Models with a random intercept for each of the respondents. The response variables were comprehensibility, intelligibility and foreign accent. The predictor variables were speaker (i.e. speaker A vs. B), context (i.e. meaningful sentences vs. nonsensical sentences) and the two-way interaction between speaker and context. In these Linear Mixed-effects Models, the attitudinal dimensions resulting from the 17 questions in task 3 were also added as predictor variables. These attitudinal dimensions were obtained with Principal Component Analysis. The results of both the Principal Component Analysis and the Linear Mixed-effects Models are shown in Section 4.2.

4. Results

4.1. Listeners' characteristics vs. general attitudes towards non-native speakers/speech

To answer our first research question (RQ1) we examined whether listeners' characteristics can predict their general attitudes towards non-native speakers and speech. We focus on the impact of the background variables "age", "educational level", "extent of contact with L2 speakers" and "political preference" on the responses to the set of statements about non-native speakers and speech.

Before we elaborate on the correlation between these background variables and the answers to the statements about non-native speakers and speech, we give an overview of (an English translation of) these statements. We calculated the mean, standard deviation, and median of the answers to these statements, and give an overview of the number of respondents (out of a total of 126) who answered with

"strongly disagree", "somewhat disagree", "neutral", "somewhat agree" and "strongly agree". These descriptive statistics are referred to below, when we discuss the link between listener's characteristics and their attitudes.

Table 2. Descriptive general attitudes towards non-native speakers/speech

	Statement in English	Mean	Standard deviation	Median	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree
LEARN_1	Newcomers who want to continue living in Flanders have to learn Dutch.	4.46	0.59	5	0	0	6	56	64
LEARN_2	I feel annoyed when migrants have not learnt Dutch after being in Flanders for many years.	4.05	0.94	3	2	9	13	59	43
LEARN_3	Many non-native speakers make an effort to learn Dutch.	3.31	0.98	4	3	25	41	44	13
POL_1	There are too many non-native speakers in Flanders.	2.98	1.04	3	10	29	51	26	10
POL_2	I believe the government spends too much money on Dutch language classes for non-native speakers.	2.41	0.87	2	17	54	42	12	1

POL_3	I think it is right that the government spends a lot of money on Dutch classes for non-native speakers.	3.72	0.89	4	1	12	30	61	22
POL_4	I think it is right that the Flemish government spends a lot of money on non-native speakers.	3.02	0.91	3	2	36	55	24	9
POL_5	I think it is a pity that non-native speakers are obliged to learn Dutch.	1.54	0.69	1	68	52	2	4	0
SLA_1	Learning Dutch is quite hard for non-native speakers.	4.02	0.89	4	1	9	15	62	39
SLA_2	It is normal for non-native speakers of Dutch to keep a foreign accent.	4.24	0.71	4	0	3	11	65	47
SLA_3	I expect non-native speakers of Dutch who intend to stay in Flanders to lose their foreign accent after one year.	1.47	0.68	1	78	39	7	2	0
SLA_4	I expect non-native speakers who intend to stay in Flanders to learn to speak Dutch flawlessly.	2.77	1.10	2	13	49	23	36	5

EFFORT_1	I find it tiring to listen to non-native speakers of Dutch.	2.14	0.91	4	30	61	23	11	1
EFFORT_2	I continue to listen to non-native speakers of Dutch, even though it requires a lot of effort to understand them.	4.21	0.74	2	1	3	9	68	45
EFFORT_3	I find it annoying to make an effort to understand non-native speakers of Dutch.	2.05	0.85	4	32	66	18	10	0
EFFORT_4	Having a conversation with non-native speakers requires a lot of focus.	3.55	0.79	2	1	14	32	73	6
STATUS_1	Dutch with a foreign accent is suitable for a teacher.	2.59	0.97	2	14	52	34	24	2
STATUS_2	It is okay for a newsreader to speak Dutch with a foreign accent.	2.48	1.06	2	23	49	28	23	3
STATUS_3	It bothers me when a shop assistant speaks Dutch with a foreign accent.	1.70	0.68	2	51	65	7	3	0
STATUS_4	I take a non-native speaker of Dutch less seriously.	1.83	0.90	5	52	54	9	11	0

SOL_1	I could be friends with a non-native speaker who doesn't speak Dutch well.	4.36	0.78	4	0	4	12	45	65
SOL_2	I trust non-native speakers of Dutch.	3.90	0.88	3	0	6	37	46	37
SOL_3	Most non-native speakers of Dutch go to great lengths to integrate into society.	3.12	0.96	3	5	32	36	49	4
SOL_4	I feel safe with non-native speakers.	3.45	0.91	3	3	13	48	48	14
SOL_5	I enjoy interacting with non-native speakers in Flanders.	3.54	0.92	2	1	12	53	38	22
EST_1	I think Dutch with a foreign accent is not attractive.	2.22	0.92	3	30	50	34	12	0
EST_2	Dutch with a foreign accent sounds nice.	3.06	0.73	3	4	17	73	31	1
INTELL_1	Non-native speakers of Dutch are often difficult to understand.	3.10	0.84	3	0	37	42	45	2
INTELL_2	Conversations with non-native speakers of Dutch often run smoothly.	2.98	0.84	3	1	41	46	36	2
INTELL_3	A conversation with a non-native speaker of Dutch often leads to misunderstandings.	2.80	0.86	3	4	48	44	29	1

INTELL_4	I have to make a huge effort to understand non-native speakers of Dutch.	2.82	0.83	2	2	51	41	32	0
VAR_1	Mistakes in Dutch bother me.	2.71	1.22	4	22	42	22	31	9
VAR_2	It is important that everyone should be able to speak Standard Dutch.	3.55	1.02	3	5	18	22	65	16
VAR_3	Dialect and substandard language may also be used in public broadcasting (radio and television).	2.99	1.14	3	12	39	20	48	7

Below we present the associations which we found between the background variables age, educational level, extent of contact with L2 speakers and political preference, on the one hand, and the 34 statements probing social attitudes, on the other hand. There was a significant effect of age on 4 statements, of educational level on 5 statements, of extent of contact with L2 speakers on 3 statements and of political preference on 18 statements. At the same time, this means we observed no significant effect of age on 30 statements, of educational level on 29 statements, of extent of contact on 31 statements and of political preference on 16 statements.

We observed a significant effect of the age of our respondents on their answers to the statements LEARN_1, LEARN_2, EST_1 and VAR_3 (cf. Table 3). Even though all our respondents generally seemed to find it important for newcomers to learn Dutch (cf. Table 2), we did observe age differences: the older the respondents, the more strongly they believe newcomers have to learn Dutch (LEARN_1) and the more they feel annoyed when migrants have not succeeded in learning Dutch, despite having lived in Flanders for many years (LEARN_2). Older respondents are also significantly less tolerant towards the use of non-standard Dutch on radio and television (VAR_3). At the same time, though, our younger respondents seem to find Dutch with a foreign accent less attractive than the older ones (EST_1).

Table 3. Statements with a significant effect of age (the intercept provides an estimate of the response variable at age 0;0). Between brackets are the 95% confidence intervals

Beta	LEARN_1	LEARN_2	EST_1	VAR_3
Intercept	2.3770 (1.4004; 3.3535)	1.8723 (0.3812; 3.3633)	2.5224 (0.9542; 4.0906)	5.2956 (3.5002; 7.0910)
AGE	-0.0180 (-0.0294; -0.0066)	-0.0178 (-0.0353; -0.0004)	0.0225 (0.0042; 0.0409)	-0.0525 (-0.0734; -0.0315)

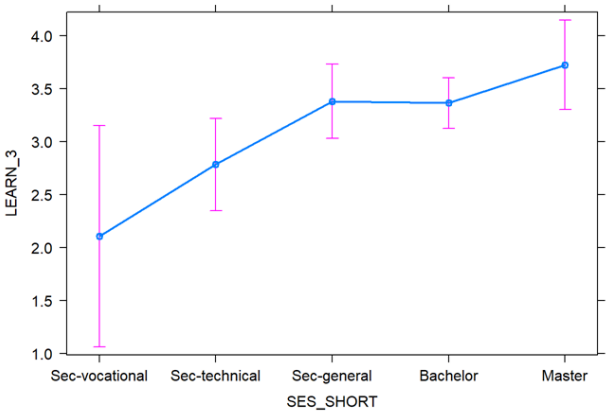
Besides age differences, we also observed a significant effect of the educational level of our respondents on their answers to the statements LEARN_3, POL_3, STATUS_4, SOL_3 and INTELL_2 (cf. Table 2). Note that secondary education: vocational schooling is the reference level of this categorical variable, so its estimated value (for each of the six statements) is represented by the intercept.

Table 4. Statements with a significant effect of educational level (compared to "secondary education: vocational schooling" as a reference point). Between brackets are the 95% confidence intervals

Beta	LEARN_3	POL_3	STATUS_4	SOL_1	SOL_4	VAR_3
Intercept	2.7395 (1.1884; 4.2906)	2.8540 (1.3672; 4.3408)	2.9568 (1.4451; 4.4685)	3.1053 (1.7992; 4.4114)	2.5064 (1.0031; 4.0097)	5.2956 (3.5002; 7.0910)
Secondary education: technical schooling	0.6770 (-0.4380; 1.7921)	1.1434 (0.0745; 2.2122)	0.8188 (-0.2680; 1.9055)	0.3669 (-0.5720; 1.3058)	0.4887 (-0.5921; 1.5694)	-0.9916 (-2.2823; 0.2991)
Secondary education: general schooling	1.2748 (0.1425; 2.4072)	1.7909 (0.7055; 2.8763)	1.5054 (0.4018; 2.6090)	0.7553 (-0.1982; 1.7087)	0.7981 (-0.2994; 1.8956)	-1.0500 (-2.3607; 0.2607)
Bachelor's degree	1.2596 (0.1951; 2.3241)	1.2745 (0.2541; 2.2949)	0.9987 (-0.0388; 2.0362)	0.7022 (-0.1942; 1.5985)	0.6797 (-0.3521; 1.7114)	-1.2142 (-2.4464; 0.0180)
Master's degree	1.6186 (0.4949; 2.7424)	1.7326 (0.6555; 2.8098)	1.0635 (-0.0317; 2.1587)	0.9562 (0.0100; 1.9025)	1.1510 (0.0619; 2.2402)	-1.3785 (-2.6793; -0.0777)

The participants with a lower educational level more strongly believed that there are too many non-native speakers of Dutch in Flanders (LEARN_3, see Figure 1) and were less positive about the government spending large sums of money on Dutch language classes for non-native speakers (POL_3). We observe a similar impact of educational level on the extent to which respondents take non-native speakers of Dutch seriously (STATUS_4), the extent to which they believe that most non-native speakers of Dutch go to great lengths to integrate into society (SOL_3) and the extent to which they have the impression conversations with non-native speakers of Dutch run smoothly (INTELL_2).

Figure 1. Responses to LEARN_3 ("Many non-native speakers make an effort to learn Dutch") in relation to educational background



In all statements, we observe less positive attitudes in respondents with a qualification in vocational and technical secondary education than in respondents with a qualification in general secondary education or with a Bachelor's or Master's degree. This does not mean that respondents with a lower educational qualification always held negative attitudes, but their attitudes were less positive than those of respondents with a higher educational qualification. For instance, most of the respondents disagreed with the statement that they take non-native speakers less seriously, but respondents with a higher educational qualification disagreed more strongly (cf. Table 2).

The background variable extent of contact with L2 speakers showed an effect with the statements SOL_1, SOL_5 and INTELL_2 (cf. Table 5). In general, most respondents seemed to (completely) agree with SOL_5 and especially SOL_1: they were quite positive about contacts with L2 speakers (cf. Table 2). In addition, we observed some differences depending on the extent of contact with L2 speakers: the

more contact respondents have with non-native speakers of Dutch, the more they feel they could be friends with them (SOL_1) and the more they enjoy interacting with them (SOL_5). They also more strongly believe conversations with non-natives often run smoothly (INTELL_2).

Table 5. Statements with a significant effect of extent of contact (compared to respondents who are "rarely or never" in contact with non-native speakers as a reference point). Between brackets are the 95% confidence intervals

Beta	EFFORT_1	SOL_1	SOL_2	SOL_5	INTELL_2
Intercept	2.1419 (0.5677; 3.7162)	3.1053 (1.7992; 4.4114)	2.3660 (0.9272; 3.8047)	2.9775 (1.5249; 4.4301)	2.3011 (0.8889; 3.7133)
Monthly contact	0.4978 (-0.1389; 1.1344)	0.3771 (-0.1511; 0.9053)	0.4445 (-0.1373; 1.0264)	0.4781 (-0.1094; 1.0655)	0.7072 (0.1361; 1.2783)
Weekly contact	0.3913 (-0.2248; 1.0074)	0.4228 (-0.0883; 0.9340)	0.3348 (-0.2283; 0.8978)	0.3089 (-0.2596; 0.8773)	0.3390 (-0.2137; 0.8917)
Daily contact	0.7716 (0.1132; 1.4300)	0.9110 (0.3647; 1.4573)	0.6864 (-0.2283; 0.8978)	1.0175 (0.4099; 1.6250)	0.7241 (0.1335; 1.3147)

The background variable which showed most significant effects is political preference. There is a significant effect of political preference on the answers of the following 18 statements: LEARN_1, LEARN_2, LEARN_3, POL_1, POL_2, POL_3, POL_4, POL_5, STATUS_2, STATUS_3, STATUS_4, SOL_2, SOL_3, SOL_4, SOL_5, EST_1, INTELL_1 and SLA_4 (cf. Table 6).

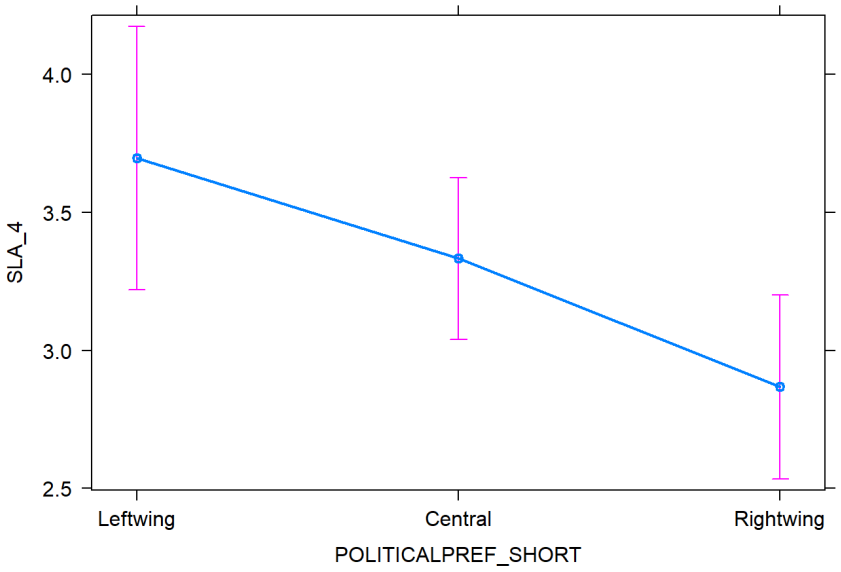
Table 6. Statements with a significant effect of political preference (compared to left-wing respondents as a reference point). Between brackets are the 95% confidence intervals

Beta	LEARN_1	LEARN_2	LEARN_3	POL_1	POL_2	POL_3
Intercept	2.3770 (1.4004; 3.3535)	1.8723 (0.3812; 3.3633)	2.7395 (1.1884; 4.2906)	3.8861 (2.2780; 5.4942)	3.2509 (1.8627; 4.6391)	2.8540 (1.3672; 4.3408)
Central	-0.2303 (-0.5203; 0.0598)	-0.2732 (-0.7161; 0.1696)	-0.5648 (-1.0255; -0.1041)	-0.5441 (-1.0217; -0.0665)	-0.7473 (-1.1596; -0.3350)	-0.1732 (-0.6148; 0.2684)
Right-wing	-0.5140 (-0.8089; -0.2191)	-0.9367 (-1.3870; -0.4865)	-0.9611 (-1.4296; -0.4927)	-1.2971 (-1.7827; -0.8114)	-0.9043 (-1.3236; -0.4851)	-0.5326 (-0.9816; -0.0836)
Beta	POL_4	POL_5	EST_1	STATUS_2	STATUS_3	STATUS_4
Intercept	4.2000 (2.8221; 5.5779)	2.1267 (0.9660; 3.2873)	2.5224 (0.9542; 4.0906)	3.2066 (1.3581; 5.0550)	3.8876 (2.7331; 5.0421)	2.9568 (1.4451; 4.4685)
Central	-0.4249 (-0.8342; -0.0157)	-0.3812 (-0.7259; -0.0365)	-0.4614 (-0.9272; 0.0044)	-0.0191 (-0.5681; 0.5300)	-0.3758 (-0.7187; -0.0329)	-0.3485 (-0.7975; 0.1005)
Right-wing	-1.1305 (-1.5466; -0.7144)	-0.6084 (-0.9589; -0.2579)	-0.6033 (-1.0769; -0.1297)	-0.6484 (-1.2067; -0.0902)	-0.5551 (-0.9037; -0.2065)	-0.6383 (-1.0948; -0.1818)
Beta	SOL_2	SOL_3	SOL_4	SOL_5	INTELL_1	SLA_4
Intercept	2.3660 (0.9272; 3.8047)	2.8390 (1.2885; 4.3895)	2.5064 (1.0031; 4.0097)	2.9775 (1.5249; 4.4301)	2.2757 (0.8271; 3.7244)	4.5772 (2.6524; 6.5020)
Central	0.1352 (-0.2921; 0.5626)	-0.2480 (-0.7085; 0.2125)	-0.2003 (-0.6468; 0.2462)	-0.1438 (-0.5753; 0.2876)	-0.2164 (-0.6466; 0.2139)	-0.3630 (-0.9347; 0.2087)
Right-wing	-0.2965 (-0.7309; 0.1380)	-0.8034 (-1.2717; -0.3352)	-0.7378 (-1.1918; -0.2838)	-0.6196 (-1.0583; -0.1810)	-0.6596 (-1.0971; -0.2221)	-0.8279 (-1.4092; -0.2466)

Even though most respondents seem to agree with the statements LEARN_1 and LEARN_2 (cf. Table 2), the right-wing respondents more strongly believe that

newcomers who want to continue living in Flanders have to learn Dutch (LEARN_1) and they feel more annoyed when migrants have not learned the language after having lived in Flanders for many years (LEARN_2). Right-wing respondents are also less convinced that many non-native speakers in Flanders make an effort to learn Dutch (LEARN_3). Moreover, they have higher expectations regarding the level of Dutch learners have to attain than left-wing respondents: they more strongly believe that newcomers should learn to speak Dutch flawlessly (SLA_4), as shown in Figure 2.

Figure 2. Responses to SLA_4 ("I expect non-native speakers who intend to stay in Flanders to learn to speak Dutch flawlessly") in relation to political preference



As is clear from Figure 2 about SLA_4, the answers of respondents who situate themselves mid-spectrum politically, lie somewhere in between left- and right-wing. This is also the case for the other statements mentioned in Table 6.

Especially the answers to the statements we categorised under "Political preference" (and more specifically statements POL_1 and POL_4) seemed to cause division between the respondents (cf. Table 2). The right-wing respondents agree more often with the statement that there are too many non-native speakers of Dutch in Flanders (POL_1). They also more strongly believe that the government spends too much money on Dutch language classes for non-native speakers (POL_2, POL_3) and they approve less often of the fact that the Flemish government spends a lot of money on non-native speakers in general (POL_4). Compared to left-wing respondents, they

agreed more strongly or disagreed less strongly with the statement "I think it is a pity that non-native speakers should be obliged to learn Dutch" (POL_5).

Even though none of our respondents were really open to the idea of a newsreader speaking with a non-native accent of Dutch (STATUS_2: M=2.48; SD=1.06), right-wing respondents were still less fond of the idea. In general, all our respondents were less bothered by the idea that a shop assistant would have a foreign accent in Dutch (STATUS_3), but we again observe differences between right-wing and left-wing respondents. The right-wing respondents also take non-native speakers of Dutch less seriously (STATUS_4).

In the same vein, right-wing respondents are less trusting of non-native speakers (SOL_2) and feel less safe with them (SOL_4). They also enjoy having contact with them less than central and left-wing respondents (SOL_5). Moreover, they are less convinced than central and left-wing respondents that most non-native speakers of Dutch go to great lengths to integrate into society (SOL_3). Finally, they more strongly believe that Dutch with a foreign accent is unattractive (EST_1) and that non-native speakers of Dutch are often difficult to understand (INTELL_1).

4.2. Listener's evaluations of non-native speakers/speech

To answer the question if listeners' evaluations of non-native speakers and non-native speech can predict the perceived foreign accentedness, comprehensibility and intelligibility of non-native speech (RQ2), we first needed to check (1) if significant differences could be observed between the two speakers in these respects and (2) if the 17 statements of the speaker/speech evaluation experiment can be grouped into dimensions.

4.2.1. Comparison of speakers and contexts

The 102 respondents in the second part of our study did not rate the two speakers significantly different in terms of comprehensibility. For intelligibility (i.e., the accuracy with which target words were transcribed by participants), however, there was a significant interaction between the two speakers and the two contexts (cf. Table 7). This is visualized in Figure 3.

Table 7. Effect of speaker and context on intelligibility (with speaker A and "CONTEXT:Real" as reference points)

	Estimate	2.5 %	97.5 %
(Intercept)	0.9020	0.8771	0.9269
SPEAKER:B	0.0283	-0.0037	0.0602
CONTEXT:Nonsense	-0.0553	-0.0873	-0.0233
SPEAKER:B x CONTEXT:Nonsense	-0.2096	-0.2548	-0.1644

Figure 3. Intelligibility speaker A vs. speaker B (percentage of correctly transcribed items) and nonsensical ("nonsense") vs. semantically meaningful ("real") sentences

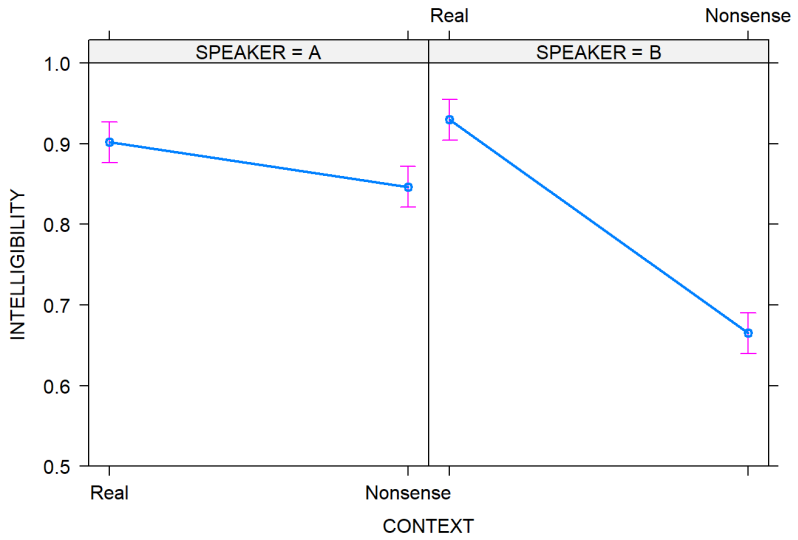


Figure 3 shows that for both speakers the intelligibility score for the semantically meaningful sentences is significantly higher than for the nonsensical sentences; in the nonsensical sentences, speaker B is significantly less intelligible than speaker A. This means that we have to study the impact of the evaluations of non-native speakers and speech on the intelligibility of these two speakers and two contexts separately.

We also observed significant differences between the two speakers in terms of strength of foreign accent: speaker B is assigned a less strong accent than speaker A, which is surprising as speaker B received a lower intelligibility score for the nonsense sentences (cf. Table 8).

Table 8. Effect of speaker on (strength of) foreign accent (with "Speaker A" as reference point)

	Estimate	2.5 %	97.5 %
(Intercept)	3.9753	3.8516	4.0989
SPEAKER:B	-0.2545	-0.4006	-0.1084

4.2.2. PCA of the speaker/speech evaluation experiment

Principal component analysis with Varimax rotation revealed that the responses to the 17 statements of the speaker/speech evaluation task correlated in five dimensions, which could be identified as "effort required from listener", "solidarity", "accent appreciation", "status" and "level of integration". Table 9 gives an overview of the dimensions and statements within these dimensions. As a result, the analyses in Sections 4.2.3, 4.2.4 and 4.2.5 are based on these five dimensions.

Table 9. PCA speaker/speech evaluation experiment

Variable	PA1: effort required from listener	PA2: solidarity	PA3: accent appreciation	PA4: status	PA5: level of integration
Understanding this speaker requires a lot of focus.	0.861				
I find it hard/difficult to listen to this speaker.	0.811				
I find it easy to understand this speaker.	0.771				
I find the language of this speaker irritating.	0.763				
I find the language of this speaker beautiful.	0.643				
I think this speaker can be trusted.		0.865			
I think this speaker is a fun person.		0.739			
I think this speaker has a lot of friends.		0.617			
I think this speaker is unemployed.		0.590			
I find this speaker's foreign accent funny.			0.855		
I find the foreign accent of this speaker cute.			0.776		
I like the foreign accent of this speaker.			0.753		
I think this speaker has a degree in higher education.				0.730	
I think this speaker holds a high position.				0.870	
I think this speaker is well integrated in Flanders.					0.833
I think this person has adapted to the social norms and customs in Flanders.					0.688
I think this speaker knows many Dutch speakers in Flanders.					0.579

4.2.3. Comprehensibility vs. speaker/speech evaluation

Although the 17 attitude questions were asked about speakers A and B separately, there were no significant differences between the two speakers in terms of the five attitude variables from the PCA. This allowed us to limit ourselves to studying the main effects of the five attitude variables on comprehensibility and foreign accent. We observed a significant effect of the attitude variables "effort required from the listener" (PA1) and "status" (PA4) on comprehensibility, as shown in Table 10.

Table 10. Effects of attitude variables PA1 ("Effort required from listener") and PA4 ("Status") on comprehensibility

	Estimate	2.5 %	97.5 %
(Intercept)	3.8088	3.6856	3.9320
PA1: effort required from the listener	0.3461	0.2284	0.4644
PA4: status	0.1913	0.0646	0.3180

The less effort it took respondents to understand the speakers (PA1) and the higher the speakers were rated on the status dimension (PA4), the higher they were also rated for comprehensibility.

4.2.4. Perceived foreign accent vs. speaker/speech evaluation experiment

In the evaluation of foreign accent, our 102 respondents gave a significantly different rating to the two speakers (cf. Section 3.2.2.2). The foreign accent of speaker A is perceived to be stronger than that of speaker B, as shown in Table 11.

Table 11. Effect of attitude variables PA1 ("Effort required from the listener"), PA2 ("Solidarity") and PA4 ("Status") on perceived foreign accent

	Estimate	2.5 %	97.5 %
(Intercept)	3.9753	3.8516	4.0989
SPEAKER:B	-0.2545	-0.4006	-0.1084
PA1: effort required from the listener	-0.2874	-0.3883	-0.1864
PA2: solidarity	0.1220	0.0202	0.2238
PA4: status	-0.1443	-0.2516	-0.0370

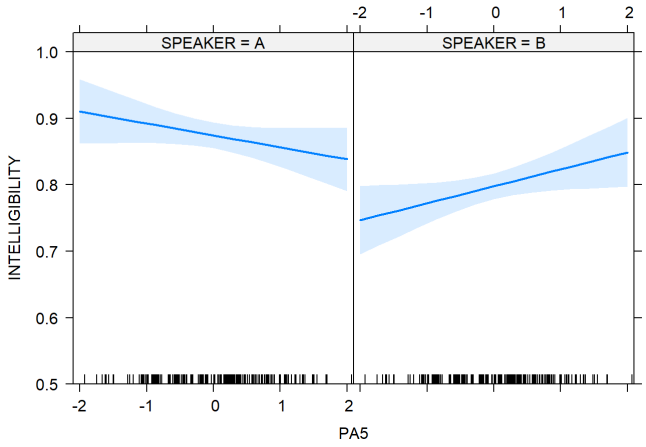
We observed a significant effect of the attitude variables "effort required from listener" (PA1), "solidarity" (PA2) and "status" (PA4) on the evaluation of foreign accent. These three effects were the same for both speakers (since none of the interaction effects of the attitude variables with speaker were statistically significant).

Foreign accent is perceived to be stronger when it requires more effort to understand the speakers (PA1) and the speakers were rated higher on the solidarity (PA2) and lower on the status dimension (PA4) when their foreign accent was perceived to be stronger.

4.2.5. Intelligibility vs. speaker/speech evaluation experiment

Concerning intelligibility, there was a significant difference between the two speakers. The difference resided in the effect of the attitude variable "level of integration" (PA5) on their intelligibility, as visualized in Figure 4. Figure 4 shows that, whereas the attitude variable appeared to adversely affect the intelligibility of speaker A, it proved to have a positive bearing on the intelligibility of speaker B. This means that listeners who believed speaker A to be better integrated in Flanders (higher score on the X-axis) had greater difficulty understanding that speaker than listeners who believed the speaker to be less well integrated. The opposite trend was observed for speaker B.

Figure 4. Effect of "level of integration" on intelligibility of speaker A and B



We also observed a significant difference between the two contexts (i.e. meaningful sentences vs. nonsensical sentences) in the effect of the attitude variable "effort required

from the listener" on the intelligibility of both speakers (cf. Figure 5; see Table 12). In other words, listeners who believed that conversations with the non-native speakers would require a lot of effort from them were better able to understand the speakers in the real sentences, but the reverse was true in the nonsensical sentences.

Figure 5. Effect of effort required from the listener on intelligibility in semantically meaningful ("real") and nonsensical ("nonsense") sentences

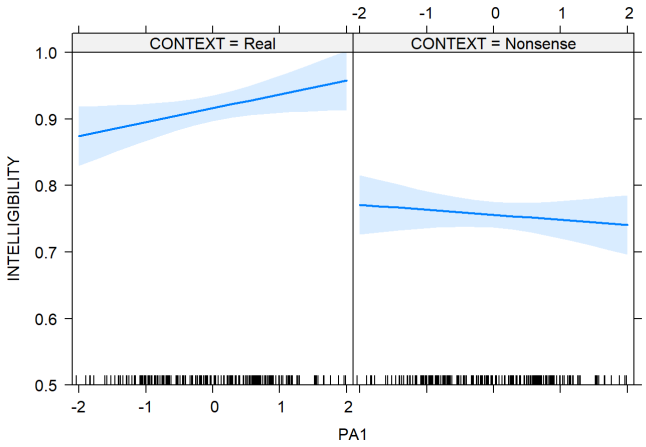


Table 12. Effect of attitude variables PA1 ("Effort required from the listener") and PA5 ("Level of integration") on intelligibility

	Estimate	2.5 %	97.5 %
(Intercept)	0.9020	0.8771	0.9269
SPEAKER:B	0.0283	-0.0037	0.0602
CONTEXT:Nonsense	-0.0553	-0.0873	-0.0233
PA5: level of integration	-0.0179	-0.0397	0.0039
PA1: effort required from the listener	0.0210	0.0011	0.0409
SPEAKER:B x CONTEXT:Nonsense	-0.2096	-0.2548	-0.1644
SPEAKER:B x PA5: level of integration	0.0435	0.0139	0.0731
CONTEXT: Nonsense x PA1	-0.0286	-0.0533	-0.0038

5. Discussion

This study had two main research aims. The first aim was to investigate whether the listener characteristics "age", "level of education", "extent of contact with L2 speakers" and "political preference" predict listeners' attitudes towards non-native speakers and non-native speech in general. To this end, 126 Dutch-speaking Belgian respondents completed a 34-item questionnaire probing their views on themes such as the effort needed to communicate with non-native speakers, the status of non-native speakers, solidarity towards non-native speakers and the effort invested by non-native speakers in learning Dutch. Previous studies aimed at determining which factors predict respondents' social attitudes tend to focus on attitudes towards migrants and refugees (e.g. Chandler & Tsai, 2001; Butkus, Maciulyte-Sniukiene & Matuzeviciute, 2016; Murray & Marx, 2013; see Section 2.2). We hypothesized that the same characteristics would help to explain people's social attitudes towards non-native speakers and their speech in general, i.e. without the respondents' knowledge whether these speakers are immigrants or refugees or neither. The results of the analysis of the questionnaire data in relation to respondents' background characteristics confirmed this hypothesis, in the sense that all four factors – "age", "level of education", "extent of contact with L2 speakers" and "political preference" – predicted the responses to a number of statements. As far as the age of the respondents is concerned, we found that the results are generally in line with Butkus et al.'s (2016) claim that more tolerant social attitudes are typically associated with youth: younger respondents were found to be generally less insistent on newcomers having to learn Dutch as soon as possible, less bothered by newcomers who have not learnt Dutch (well) yet and more tolerant towards variation in Dutch (e.g. towards native but substandard accents in the media (radio and television) than older respondents. The educational level of the respondents also proved to play a role, in that we observed more positive attitudes towards non-native speakers and their speech in respondents with a higher level of education. For instance, these respondents generally did not agree that there are too many non-native speakers of Dutch in Flanders, they thought it is right that Dutch language classes are government-funded and agreed that most non-native speakers of Dutch go to great lengths to integrate into society. These findings are in line with previous studies on social attitudes towards immigrants or refugees. Those studies hypothesize that the absence of economic competition with newcomers and the more diverse and cosmopolitan networks of people with a higher educational level foster tolerance (Butkus et al., 2016; Chandler & Tsai, 2001). Previous findings are also confirmed for the variable "extent of contact with L2 speakers": respondents who reported to have more contact also felt more strongly that they could be friends with non-native speakers and believed conversations with non-native speakers to generally run smoothly (cf. the Contact Theory mentioned in Section 2.2, Butkus et al., 2016). Finally, we observed a

significant correlation between political preference and the answers to 18 statements including statements related to the status of and solidarity with non-native speakers and the aesthetic value of accented speech. For instance, respondents who described their political ideology as right-wing have higher expectations of the level of Dutch that learners have to attain, are significantly less convinced that many non-native speakers in Flanders make an effort to learn Dutch, trust non-native speakers significantly less and feel significantly less safe with them. The answers of the respondents who characterize their political ideology as "centrist" typically lie somewhere in between left- and right-wing. These findings are in line with earlier studies on the correlation between political stance and social attitudes towards immigrants, which have shown that people with right-wing or conservative political ideologies tend to adopt a more negative stance towards newcomers (Chandler & Tsai, 2001).

In conclusion, a number of statements led to responses which could be predicted by the four listener characteristics discussed. However, it is also important to mention that for many of the statements, no effect of these characteristics was observed. For instance, we found that the variable "age" predicted responses to 4 statements, meaning that it had no predictive power for responses to the remaining 30 statements. The same holds true for educational level (significant predictor for 5/34 statements) and extent of contact with L2 speakers (significant predictor for 3/34 statements). In other words, the impact of "age", "level of education" and "contact with L2 speakers" is certainly present, but it does not predict the responses to the majority of statements viz. non-native speakers and non-native speech. The effect of political preference proved to be considerably stronger: respondents' self-confessed political ideology predicted their responses to 18 out of 34 statements. The strong predictive power of political preference is not surprising, as it has been observed that migration policy is one of the most important electoral motives in Belgium and voting behaviour is strongly influenced by attitudes towards immigrants and, by extension, towards non-native speakers (Deschouwer et al., 2010). We should, however, also add a cautionary note here. The participants were recruited from the researchers' social networks through social media (e.g. Facebook profiles of Master students involved in the recruitment procedure). This means that the sample may not be representative of the larger population. For instance, participants with a Bachelor's, Master's or PhD degree seem to be overrepresented. The sampling method may thus, to a certain extent, skew the results.

In the second part of the study, we establish the link between people's social attitudes towards non-native speakers and the intelligibility of non-native speech. Specifically, the second aim of the study was to investigate whether listeners' evaluations of non-native speakers and their speech predict their responses with regard

to intelligibility, comprehensibility and (foreign) accentedness of these speakers. To answer this question, 122 respondents performed speaker and speech evaluation tasks with samples of L2 Dutch speech produced by two native speakers of Mandarin Chinese with a high proficiency in Dutch. The results of these speaker/speech evaluation experiments were analysed in relation to respondents' performance on a transcription task and their ratings of the speakers' comprehensibility and accentedness. A comparison between the two speakers showed that listeners rated the speakers significantly differently for foreign accentedness, but not for comprehensibility. As far as intelligibility is concerned, the results of the transcription task showed that the intelligibility scores for the semantically meaningful sentences were significantly higher than for the nonsensical sentences. This suggests, as corroborated by Kang et al. (2018) (see Section 2.1.), that listeners do use available semantic context when performing intelligibility tasks. The speaker who was rated as having a stronger (foreign) accent, actually turned out to be more intelligible on the items in the nonsensical sentences than the speaker with the milder accent. These results are in line with earlier research by, among others, Munro & Derwing (1995a) and Derwing & Munro (1997), which shows that intelligibility, comprehensibility and accentedness are indeed different and partly independent constructs.

The analysis further reveals that the responses to the statements of the speaker/speech evaluation task correlated in five dimensions, which could be identified as "effort required from listener", "solidarity", "accent appreciation", "status" and "level of integration". We hypothesized that listeners' responses in the speaker/speech evaluation experiment would have an effect on the intelligibility, comprehensibility and accentedness of these same speakers. Since previous research shows mixed results, we predicted that this effect is a subtle one (see Section 3.1.). As far as comprehensibility and foreign accentedness are concerned, we did find a significant correlation between these variables and a number of attitudinal dimensions. For instance, respondents who rated the speakers high for comprehensibility also rated them higher on the status dimension. Similarly, respondents who considered the speakers to have a stronger foreign accent also reported it took more effort to understand them, and they rated them lower on the status dimension. Finally, we observed no significant correlations between any of the five attitude dimensions and intelligibility scores. We only found that the attitude variable "level of integration" impacted the two speakers differently. We also observed a significant difference between the meaningful and nonsensical sentences in the effect of the attitude variable "effort required from the listener" on the intelligibility of both speakers. These findings suggest that the link between listeners' attitudes and their scores on a transcription task measuring speakers' intelligibility is not straightforward, i.e., we can, for instance, not conclude that listeners holding, negative attitudes towards non-native speakers understand non-native speech less well

than listeners with a more positive attitude. The results indicate that the relationship between listeners' attitudes and the intelligibility of non-native speakers is subject to both speaker and task effects. It is, for instance, possible that an effect of attitudes on the success of communication with non-native speakers can only be observed when direct and personal interaction with non-native speakers is part of the experimental design, as in Lindeman (2002; cf. Section 2.2). Whether communication is successful or not, clearly depends on much more than on the intelligibility of individual words. Segmental as well as prosodic aspects play a role (Caspers & Horloza, 2012) and, above all, conversations do not take place in a situational vacuum, but are embedded in a social and cultural context (Rajadurai, 2007). In the present study, the L2 Dutch speech samples of the L1 Mandarin Chinese speakers were presented only aurally, without a face attached to them. This may have diminished the influence of social attitudes on the perception of the non-native speech.

6. Conclusions and suggestions for further research

As many Western societies become increasingly diverse due to mobility and migration, the number of people communicating with non-native speakers is rapidly growing and tolerance as a social attitude is becoming more and more imperative (Hooghe, Meeusen & Quintelier, 2013). In this study, we set out to explore people's background characteristics affecting social attitudes towards non-native speakers, as well as the effect of these attitudes on the intelligibility, comprehensibility and accentedness of non-native speech. The results of a survey confirmed that the listeners' age, level of education, extent of contact with L2 speakers and, above all, their political preference predicted their social attitudes. Younger, highly educated people who had more contact with non-native speakers and situated themselves more towards the left of the political spectrum, had more lenient attitudes towards non-native speakers and their speech. The result of the experimental part of the study confirmed that there was a link between people's evaluations of non-native speakers and their speech and measures of comprehensibility and accentedness. Listeners who rated the speakers high on the status dimension also reported they found them more comprehensible. Conversely, listeners who rated the speakers low on the status dimension reported hearing a stronger foreign accent and experiencing greater difficulty understanding the speakers. While it is open to debate whether the listeners' attitudes affected their comprehensibility and accentedness judgments or vice versa, the results show that there is a correlation between attitudes on the one hand and comprehensibility and accentedness judgments on the other hand. We should obviously be careful when generalizing the results. In the current study, both non-native speakers of Dutch in Part 2 of the research are native speakers of Mandarin Chinese. It is well-known that

attitudes towards native speakers of different native languages differ, depending on the status and prestige attached to certain ethnic backgrounds. In that sense, Jaspers (2009:19) argues that two types of multilingualism can be distinguished, which he calls "prestigious" and "plebeian" multilingualism. As our study included only Chinese-accented speech samples, we cannot generalize our results to include all non-native speakers and it is possible that speech samples of speakers with different native languages (e.g. French or Arabic) would have yielded different results.

The study has implications for language teaching and assessment of language proficiency. Specifically, the relationship between attitudes and comprehensibility and accentedness judgements is interesting in light of second language testing procedures, in which speaking proficiency is assessed through scales using "intelligibility", "comprehensibility" and "accentedness" as central concepts. Given the current centrality of these concepts in language assessment, it is crucial to realize that judgements may be importantly influenced by the listener's or test taker's social attitudes. Test takers should be aware of the potential effects of their personal stance towards non-native speakers and their speech on the extent to which non-native speaking students are considered to be comprehensible and their speech accented. While such biases cannot be altogether ruled out, awareness of their possible presence may help to mitigate the effect. For that reason, it would be beneficial to include awareness raising of the effects of listeners' backgrounds and social attitudes in teacher training programmes.

Finally, further research on intelligibility is needed, using longer stretches of contextualized speech to be presented to listeners or looking at intelligibility and comprehensibility ratings in real interactions between native and non-native speakers. As noted above, intelligibility was assessed through a transcription task involving isolated words and nonwords and comprehensibility and accentedness ratings were based on relatively short stretches of speech. Future studies in which participants interact with each other in tasks would help us to understand whether listeners' social attitudes towards non-native speakers and their speech can also be linked to comprehensibility and accentedness judgements when real interaction with non-native speakers takes place.

7. Acknowledgements

We are grateful to Justine De Jaeger and Merel Beeckman for their help in collecting the data for this study.

8. References

Babel, M. & Russell, J. (2015) Expectations and speech intelligibility. *Journal of the Acoustical Society of America,* 137(5), 2823-2833.

Beinhoff, B. (2014) Perceiving intelligibility and accentedness in non-native speech: a look at proficiency levels. *Concordia Working Papers in Applied Linguistics,* 5, 58-72.

Bent, T. & Bradlow, A.R. (2003) The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America,* 114, 1600-1610.

Best, C.T. (1995) A Direct Realist View of Cross-Language Speech Perception. In W. Strange (ed) *Speech Perception and linguistic experience: Issues in cross-language research.* (pp.171-204). Timonium, MD: York Press.

Best, C.T. & Tyler, M.D. (2007) Non-native and second-language speech perception: Commonalities and complementarities. In M.J. Munro & O.-S. Bohn (eds) *Second language speech learning: The role of language experience in speech perception and production*. (pp. 13-34). Amsterdam: John Benjamins.

Bohn, O.S. & Munro, M.J. (eds) (2007) *Language experience in second language speech learning: In honor of James Emil Flege*. Amsterdam: John Benjamins.

Bongaerts, T., Mennen, S. & van der Slik, F. (2000) Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a Second Language. *Studia Linguistica,* 54(2), 298-308.

Burda, A.N., Casey, A.M., Foster, T.R., Pilkingston, A.K. & Reppe, E.A. (2006) Effects of accent and age on the transcription of medically related utterances: A pilot study. *Communications Disorders Quarterly,* 27, 110-116.

Butkus, M., Maciulyte-Sniukiene, A. & Matuzeviciute, K. (2016) Socio-demographic factors influencing attitude towards refugees: An analysis of data from European social survey, *16th International Scientific Conference on Globalization and its Socio-Economic Consequences*, University of Zilina, 5th-6th October 2016, 286-294.

Card, D., Dustmann, C. & Preston, I. (2005) Understanding attitudes towards immigration: the migration and minority module of the European Social Survey, Discussion Paper Series No. 03/05, Centre for Research and Analysis of Migration, University College London, available at https://discovery.ucl.ac.uk/id/eprint/14315/1/14315.pdf., last accessed 26.5.2021.

Caspers, J. & Horłoza, K. (2012) Intelligibility of non-native produced Dutch words: Interaction between segmental and suprasegmental errors. *Phonetica*, 69, 94-107.

Chandler, C.R. & Tsai, Y. (2001) Social factors influencing immigration attitudes. *The Social Science Journal* 38, 177-188.

Council of Europe. (2001) *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.

Davies, A. (2013) *Native speakers and native users: loss and gain*. Cambridge: Cambridge University Press.

Delarue, S. (2016) *Bridging the policy-practice gap: How Flemish teachers' standard language perceptions navigate between monovarietal policy and multivarietal practice*. Unpublished Doctoral dissertation. Linguistics Department, Ghent University.

Derwing, T.M. & Munro, M.J. (1997) Accent, intelligibility and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition,* 19(1), 1-16.

Deschouwer, K., Delwit, P., Hooghe, M. & Walgrave, S. (eds) (2010) *Les Voix du Peuple*. Brussels: Editions de l'Université de Bruxelles.

Eisenstein, M. & Verdi, G. (1985) The intelligibility of social dialects for working-class adult learners of English. *Language Learning,* 35(2), 287-298.

Flege, J.E. (1995) Second-language speech learning: Theory, findings, and problems. In W. Strange (ed.) *Speech Perception and linguistic experience: Issues in cross-language research.* (pp. 229-273). Timonium, MD: York Press.

___. (2005) Misinterpretation, bias and resistance to change: The case of the Lingua Franca Core. In K. Dziubalska & J. Przedlacka (eds) *English Pronunciation Models: A changing scene.* (pp. 199-212). Bern: Peter Lang

Gallardo del Puerto, F., García Lecumberri, M. & Gómez Lacabex, E. (2015) The assessment of foreign accent and its communicative effects by naïve native judges vs. experienced non-native judges. *International Journal of Applied Linguistics,* 25(3), 202-224.

Grondelaers, S. & Speelman, D. (2013) Can speaker evaluation return private attitudes towards stigmatised varieties? Evidence from emergent standardisation in Belgian Dutch. In T. Kristiansen & S. Grondelaers (Eds.), *Language (de)standardisation in Late Modern Europe: experimental studies*. Oslo: Novus Forlag, 171-191.

Grondelaers, S. & van Gent, P. (2019) How "deep" is Dynamism? Revisiting the evaluation of Moroccan-flavored Netherlandic Dutch. *Linguistics Vanguard,* 5(1).

Hainmueller, J. & Hiscox, M.J. (2007) Educated preferences: Explaining attitudes towards immigration in Europe. *International Organization,* 61(2), 399-442.

Hayes-Harb, R., Smith, B.L., Bent, T. & Bradlow, A.R. (2008) The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics,* 36(4), 664-679.

Hooghe, M., Meeusen, C. & Quintelier, E. (2013) The impact of education and intergroup friendship on the development of ethnocentrism. A latent growth curve model analysis of a five-year panel study among Belgian late adolescents. *European Sociological Review,* 29(6), 1109-1121.

Jaspers, J. (2009) *De klank van de stad: stedelijke meertaligheid en interculturele communicatie.* Leuven: Acco.

Jenkins, J. (2000) *The Phonology of English as an International Language.* Oxford: Oxford University Press.

Kachru, Y. & Smith, L.E. (2008) *Cultures, Contexts, and World Englishes.* New York: Routledge.

Kang, O., Thompson, R.I. & Moran, M. (2018) Empirical Approaches to Measuring the Intelligibility of Different Varieties of English in Predicting Listener Comprehension. *Language Learning,* 68 (1), 115-146.

Kennedy, S. & Trofimovich, P. (2019) Comprehensibility: A useful tool to explore listener understanding. *The Canadian Modern Language Review / La revue canadienne des langues vivantes,* 75(4), 275-284.

Kent, D.R. & Miolo, G. (1994) The intelligibility of children's speech. *American Journal of Speech-Language Pathology,* 3, 81-95.

Latour, B., Van Hout, R. & Grondelaers, S. (2012) De schoonheid van taal. Hoe wezenlijk is het oordeel mooi in taalattitudes? *Taal en Tongval,* 64(2), 243-261.

Levis, J. & LeVelle, K. (eds) (2010) *Pronunciation and intelligibility: Issues in research and practice, Proceedings of the 2nd Annual Pronunciation in Language Learning and Teaching Conference,* 10-11 September 2010, Iowa State University.

Lindeman, S. (2002) Listening with an attitude: a model of native-speaker comprehension of non-native speakers in the United States. *Language in Society,* 31, 419-441.

Lybaert, C. (2017) A direct discourse-based approach to the study of language attitudes: The case of tussentaal in Flanders. *Language Sciences,* 59, 93-116.

Major, R. (2001) *Foreign Accent: The Ontogeny and Phylogeny of Non-Native Speech.* New Jersey/London: Lawrence Erlbaum Associates.

Mattys, S.L., Davis, M.H., Bradlow, A.R. & Scot, S.K. (2012) Speech recognition in adverse conditions: A review. *Language & Cognitive processes*, 27, 953-978.

Mayda, A.M. (2006) Who is against immigration? A cross-country investigation of individual attitudes towards immigrants. *Review of Economics and Statistics,* 88(3), 510-530.

Morley, J. (1991) The pronunciation component in teaching English to speakers of other languages, *TESOL Quarterly,* 25, 481-520.

Munro, M.J. (2008) Foreign accent and speech intelligibility. In H. Edwards, J.G. & M.L. Zampini (eds) *Phonology and Second Language Acquisition.* (pp. 193-218) Amsterdam: John Benjamins.

Munro, M.J. & T.M. Derwing. (1995a) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning,* 45, 73–97.

____. (1995b) Processing time, accent, and comprehensibility in the perception of foreign-accented speech. *Language and Speech*, 38, 289-306.

Murray, K.E. & Marx, D.M. (2013) Attitudes toward unauthorized immigrants, authorized immigrants and refugees. *Cultural Diversity and Ethnic Minority Psychology,* 19(3), 332-341.

Noppe, J., Vanweddingen, M., Doyen, G., Stuyck, K., Feys, S. & Buysschaert, P. (2018) *Vlaamse Migratie- en Integratiemonitor 2018*. Brussel: Agentschap Binnenlands Bestuur (available at https://www.statistiekvlaanderen.be/nl/vlaamse-migratie-en-integratiemonitor , accessed on 26.5.2021).

Rajadurai, J. (2007) Intelligiblity studies: a consideration of empirical and ideological issues, in Forum: Ideology and intelligibility. *World Englishes,* 26(1), 87-98.

Rubin, D.M. (1992) Nonlanguage factors affecting undergraduates' judgements of non-native English-speaking teaching assistants. *International Journal of Intercultural Relations,* 14(3), 337-353.

Saito, K., Trofimovich, P. & Isaacs, T. (2017) Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics,* 38(4), 439-462.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V. & Ilkan, M. (2019) How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of L1 profiles, L2 proficiency, age, experience, familiarity and metacognition. *Studies in Second Language Acquisition*, 41(5), 1133-1149.

Semyonov, M., Raijman, R. & Gorodzeisky, A. (2006) The rise of anti-foreigner sentiment in European societies, 1988-2000. *American Sociological Review,* 71(3), 426-449.

Sheppard, B.E., Elliott, N.C. & Baese-Berk, M.M. (2017) Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes,* 26, 42-51.

Taylor Reid, K., Trofimovich, P. & O'Brien, M.G. (2019) Social attitudes and speech ratings. Effects of positive and negative bias on multiage listeners' judgments of second language speech. *Studies in Second Language Acquisition,* 41, 419-442.

Winke, P., Gass, S. & Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing,* 30(2), 231-252.

Appendices

Appendix 1: Semantically meaningful and nonsensical sentences

	Dutch sentence (omitted word in bold)	English translation
Semantically meaningful sentences		
Test sentences	*Mijn kleine zus **ziet** er goed uit.*	My little sister is **looking** good.
	*De lieve prinses **heeft** een lange vlecht.*	The sweet princess **has** a long plait.
Speaker 1	*De nieuwe koper **biedt** een mooi bedrag.*	The new buyer **offers** a good price.
	*De oude man **bidt** in de kerk.*	The old man is **praying** in church.
	*Mijn rode **step** gaat toch wat snel.*	My red **scooter** goes a little fast.
	*De groene **peer** ligt in die mand.*	The green **pear** is in that basket.
	*De grijze **rat** kruipt in het hol.*	The grey **rat** is crawling into the hole.
	*Die wijze **raad** vergeet hij nooit meer.*	This wise piece of **advice** he'll never forget.
Speaker 2	*De kwade **stier** kijkt naar de lap.*	The angry **bull** is looking at the rag.
	*Het witte **schip** vaart naar de zee.*	The white **ship** is sailing to the sea.
	*De stoute jongen **vecht** met zijn broer.*	The naughty boy is **fighting** with his brother.
	*De boze vrouw **veegt** met haar bezem.*	The angry woman is **sweeping** with her broom.
	*De sterke boer **kapt** met een bijl.*	The strong farmer is **hacking** with an axe.

	*De snelle loper **kaapt** de prijs weg.*	The fast runner **takes away** the prize.
Nonsensical sentences		
Test sentences	*De zachte broek **bidt** op het bad.*	The soft trousers are **praying** on the bath.
	*De slimme **zak** voelt naar meer graan.*	The clever **bag** feels after more grain.
Speaker 1	*Mijn witte boek **piekt** naar de zee.*	My white book is **peaking** into the sea.
	*De zware jas **pikt** op zijn rug.*	The black coat is **pecking** on his back.
	*De kleine rand **heft** onder het zand.*	The small edge **lifts** under the sand.
	*Die kleine koek **heeft** veel dure spullen.*	That small biscuit **has** much expensive stuff.
	*De sterke **kaas** heeft met een bijl.*	That strong **cheese** has with an axe.
	*De snelle **kat** kijkt de prijs weg.*	That quick **cat** looks away the prize.
Speaker 2	*De rode baas **ziet** toch wat snel.*	The red boss **looks** a little quick.
	*De groene reus **zit** in die mand.*	The green giant is **sitting** in that basket.
	*De oude **zeep** zoekt in de kerk.*	The old **soap** is searching in the church.
	*Zijn grote **schep** prijst het hooi op.*	His large **shovel** is praising the hay.
	*De lieve **zaak** plakt een lange vlecht.*	The sweet **case** is sticking a long plait.
	*De slimme **zak** voelt naar meer graan.*	The clever **bag** feels for more grain.

Appendix 2: Comprehensibility and accentedness judgment task: semi-spontaneous speech

Speaker 1:

"Mijn droomreis is zeker Japan. Ik wil heel graag [euh] met mijn man gaan. [euh] Japan is een van de weinig land ter wereld met combinatie van [euh] cultuur traditie maar ook modern aspect." (18 seconds)

Speaker 2:

"[Euh] wij hebben 4 dagen euh lessen genomen [euh] met één instructor van de duikencenter [euh] en wij hebben veel [euh] vissen en dieren gezien in de zee. Ik ben echt [euh] heel blij dat [euh] ik daar kan zijn." (21 seconds)

Appendix 3: Speaker/speech evaluation task: read speech

In Parijs staat een nieuw kunstwerk. De Amerikaanse kunstenaar maakt een grote hand die tulpen vasthoudt. Daarmee wil hij de slachtoffers eren van de aanslagen op 13 november 2015. Toen stierven er in Parijs 130 mensen. (24 seconds for speaker 1; 22 seconds for speaker 2)

U-shaped Trajectories in an L2 context: Evidence from the acquisition of verb morphology

Stefan Williams
Department of Language and Literature
Norwegian University of Science and Technology
stefan.williams@ntnu.no

Pedro Guijarro-Fuentes
Department of Spanish, Modern, and Classical Languages
University of the Balearic Islands
p.guijarro@uib.es

Mila Vulchanova
Department of Language and Literature
Norwegian University of Science and Technology
mila.vulchanova@ntnu.no

Abstract

This study explores U-shaped behaviour in the acquisition of irregular verb morphology across three groups of Norwegian L2 learners of English. This phenomenon is especially interesting due to its significance for the organization and division between the mental lexicon and grammar. We hypothesized that if U-shaped behaviour was observable, then we would find significant differences in participants' performance accuracy levels in conjunction with overregularization errors. We report results on the acquisition of irregular verb morphology, in addition to mean reaction times on different types of responses (accurate responses and overregularized ones). The final analysis includes data from participants from the 8th grade (N=17), 9th grade (N=19), and 10th grade (N=15). We report results of the acquisition of irregular verbs, in addition to reaction latencies. Participants responded to an elicitation task to test performance on 40 items. The results are consistent with the later stages of U-shaped learning. We found an increase in overall accuracy co-occurring with a decrease in overregularization errors. We propose that the existence of U-shaped behaviour in the L2 is indicative of a general underlying input-driven learning pattern, and that this process is an integral part of acquiring knowledge upon exposure to irregularities in a productive paradigm.

Keywords: U-Shaped learning; irregular verb morphology; English, Norwegian; Second Language Acquisition.

Resumen

Este estudio explora el comportamiento en forma de U en la adquisición de morfología de verbos irregulares en tres grupos de estudiantes noruegos L2 de inglés. Este fenómeno es especialmente interesante por su importancia para la organización y división entre el léxico mental y la gramática. Planteamos la hipótesis de que si el comportamiento en forma de U fuera observable, entonces encontraríamos diferencias significativas en los niveles de precisión de los participantes junto con errores de sobreregularización. Este estudio informa de los resultados sobre la adquisición de morfología verbal irregular, además de tiempos medios de reacción en diferentes tipos de respuestas (respuestas precisas y sobreregularizadas). El análisis final incluye datos de aprendices del octavo grado (N = 17), noveno grado (N = 19) y décimo grado (N = 15). Informamos de los resultados de la adquisición de verbos irregulares, además de los tiempos de reacción. Los participantes respondieron a una tarea de elicitación compuesta de 40 ítems. Los resultados son consistentes con las últimas etapas del aprendizaje en forma de U. Encontramos un aumento en la precisión general que concurre con una disminución en los errores de sobreregularización. Proponemos que la existencia de un comportamiento en forma de U en la L2 es indicativo de un patrón de aprendizaje impulsado por modelos subyacentes generales, y que este proceso es una parte integral de la adquisición del conocimiento al exponerse a irregularidades en un paradigma productivo.

Palabras clave: aprendizaje en forma de U; morfología de verbos irregulares; inglés; noruego; adquisición de una segunda lengua.

1. Introduction

Phenomena such as grammatical errors and overregularization have been a topic of interest in the field of language acquisition, especially in relation to verb inflectional morphology. As pointed out by Marcus et al. (1992:1), overregularization of irregular verbs, such as "comed" instead of "came" for instance, has been observed for as long as language development has been a topic of study.

Closer studies of overregularization errors have revealed an interesting developmental curve, first reported by Ervin & Miller (1963; see also Cazden, 1968), and currently referred to as the *U-shaped developmental curve*. Children with English as their L1 produce correct irregular forms of verbs, if they mark the past tense at all, up

until their third year. However, around the age of three, they begin producing errors and overregularize irregular verbs that they previously have been able to mark correctly. Around the same time as they start producing overregularization errors, children begin acquiring the rule for marking regular past tense verbs (Pinker, 1998).

The U-shaped developmental curve has been extensively studied in first language acquisition research. While U-shaped behaviour has also been observed in the domain of second languages, few studies address this phenomenon experimentally in relation to grammar acquisition in an L2 context. This is an exploratory study seeking to attest the existence of U-shaped learning patterns in acquiring English L2 strong verb morphology within three L2 participant groups at different levels of proficiency. In particular, we were interested in the participants' performance accuracy, as well as the extent to which they tend to overregularize irregular verbs. Thus, if we find evidence of U-shaped development during L2 acquisition in the same domain it has been attested in the L1, the questions are: what is the source of this pattern, and what are the underlying mechanisms supporting it in both L1 and L2 learning (Vulchanova, Foyn, Nilsen & Sigmundsson, 2014). Such findings can shed light on how second languages are acquired in comparison to the first language. Moreover, our understanding of language development has consequences for how we approach and assess language learning in a practical sense. In a language learning environment, and for the purposes of instruction, errors might be used as a litmus to indicate what aspects need more attention. Understanding the developmental trends and which cognitive learning mechanisms are at work might help us better assess and evaluate how learners develop their second language, thereby improving the quality of second language instruction.

In exploring U-shaped development during second language grammar acquisition, an elicitation task containing 40 irregular verbs based on Berko's (1958) seminal Wug-test was designed to elicit the learners' knowledge of irregular verb morphology. Data were collected via the online survey platform *SurveyGizmo*. Norwegian learners with English as their L2 were recruited as the focus group for this study. A cross-sectional design was chosen, recruiting learners from three different school grades: the 8th grade (N=20), 9th grade (N=24), and 10th grade (N=23), respectively.

The present study is structured as follows: Section 1 reviews the literature on U-shaped trajectories in language learning and second language acquisition, while also presenting theoretical models to account for the learning curve. Section 2 describes the methodology of the study, whereas Section 3 reports results divided by group, category, and individual. In Section 4, our results are discussed and interpreted. Finally, Section 5 discusses limitations of the study and presents the main conclusions before suggesting possible future research on U-shaped development in second language acquisition.

1.1. U-shaped trajectories during language learning and other domains

Initial studies on U-shaped development began, among others, with Cazden's (1968; see also Ervin & Miller, 1963) longitudinal study of child L1 production. Cazden (1968) observed how children exhibited a period of correct performance when producing the irregular past tense verbs prior to overgeneralization errors. Overregularization of irregular nouns was also observed, whereby the participants applied the productive rule to an irregular stem resulting in forms, such as "feets" instead of "feet". Moreover, there seemed to be individual differences in error rates among participants. The same pattern of performance was later supported by a larger study utilizing data from spontaneous language produced by 83 children (Marcus et al., 1992). In addition to affirming the U-shaped learning curve, this study gave further insight into the acquisition of rule-like behaviour relating to past tense morphology.

U-shaped development during L1 acquisition of verb morphology can be summarized as follows (Pinker, 1991; Pinker & Prince, 1988; Marcus et al., 1992). Children exhibit a period of correct performance when producing irregular verbs, if they mark the past tense at all. Lasting from the age of two and into early school years, errors are produced in the form of overgeneralizing the regular inflection of the past tense, resulting in the overregularization of irregular verbs that were previously correctly inflected. However, with time, learners revert back and start marking the irregular verbs correctly again. This dip in performance (i.e., U-shape) seems to affect most irregular verbs. The phenomenon co-occurs with the acquisition of the rule of regular past tense inflection of verbs. U-shaped trajectories have been documented in other domains of first language acquisition, such as causatives (Bowerman, 1982) and pronouns (Karmiloff-Smith, 1986). This is consistent with the three-phase model of language development proposed by Karmiloff-Smith (1986), as well as with Tomasello's (1992) verb-island hypothesis, and more generally with usage-based models of language development. In these models, children's early grammars are input-driven (exemplars/item-specific), then representation-driven and top-down, and finally combining both aspects, and adult-like.

U-shaped learning is not only limited to language acquisition. Carlucci & Case (2013) report that U-shaped learning occurs in a variety of domains of child development, such as understanding temperature, weight conservation, object permanence, and facial recognition. In questioning whether U-shaped learning is logically necessary or not in relation to some formal learning tasks, Carlucci & Case (2013: 81) conclude that:

> [...] the general picture that emerges from the so-far known results [...] is that U-shaped behavior is unavoidable for full learning power in the context of a number of parametrized models of learning featuring a number of cognitively motived constraints. The results

> might be taken as suggestive of the fact that humans might exhibit U-shaped and other nonmonotonic learning patterns [...].

Marcovitch & Lewkowicz (2004) address the question of whether U-shaped developmental trajectories are more than an interesting artefact of developmental processes. They hypothesise U-shaped curves are in fact a hallmark of such processes. Drawing parallels between language acquisition and ontological adaptions, as proposed by Openheim (1981, cited in Marcovitch & Lewkowicz, 2004), they discuss how this kind of behavioural regression is central to the developmental process itself. Thus, the question arises whether U-shaped learning might indeed be a universal underlying, input-driven learning pattern in human cognitive development.

U-shaped development has played an important part in explaining different models of how the language faculty operates. It has especially been a topic of discussion for single-mechanism models of the mental faculty, such as connectionist models, and dual-mechanism models, such as the dual-mechanism hypothesis (Pinker, 1998).

1.2. U-shaped learning and the dual-mechanism hypothesis

U-shaped learning as a phenomenon has been a point of discussion on the controversial topic of the psychological reality of the organization of language knowledge in the brain, how the mental lexicon is organized and how it operates (cf. Rumelhart & McClelland, 1986; Pinker & Prince, 1988; Plunkett & Marchman, 1991; Marcus et al., 1992; Pinker 1998; Pinker & Ullman, 2002; McClelland & Patterson, 2002). The traditional account distinguishes between rote-memory and a productive system (i.e., two distinct psychological systems) (as described by Pinker & Prince, 1991). In contrast, the connectionist account seeks to make memory more powerful (Pinker, 1998) and to model language based on a neural perspective (Jackendoff, 2002). Connectionist theory maintains that there are in fact no rules and that language is the result of pattern associations and spreading activation throughout networks (cf. Rumelhart & McClelland, 1986; Plunkett & Marchman, 1991; Ellis, 2003). Importantly, concepts are acquired through iterated exposure that results in increase/decrease in connection strengths among units. Patterns of association/connection constitute knowledge representation and changes to these patterns constitute learning. As an alternative to both these theories, and combining aspects of both, the dual-mechanism theory maintains that while language relies on two modules, the mental lexicon, as a form of memory, has associative properties (Pinker, 1991).

Similar to a traditional rote-&-rule theory (e.g. Chomsky & Halle, 1968), the dual-mechanism approach suggests that there are two modules which interact with

each other to form language. First is the memory component which comprises the mental lexicon, and stores words in the form of either stems or the full irregular form. The second module is the rule-capturing component which governs the productive rules of language, and the regularities underlying grammar/language structure. This component computes the regularities on-line, utilizing the stored stems from the mental lexicon. (1) is a proposal of what such a rule could look like in relation to the productive rule of the past tense:

(1) V_{past} --> V_{stem} + *d*
(Pinker, 1998: 5)

In the case of an irregular lexical entry, however, two forms would be stored: the base form and the irregular past tense form. Additionally, there would be grammatical tags which capture the function of the word, as shown in (2).

(2) V V
Bring ~~ $Brought_{past}$
(Pinker, 1998: 6)

When English is taken into consideration, such an organization seems reasonable. Regular verbs are an open-ended class of verbs inflecting within a regular and productive paradigm. Irregular verbs, on the other hand, comprise a relatively small number of verbs (ca. 180), which inflect in idiosyncratic ways (Pinker, 1998: 5). Attempting to capture semi-regularities observed among irregular verbs, Pinker theorizes that associative memory links features to features. Thus, items which share features are assumed to be superimposed in their memory representations, reinforcing shared patterns. New items, similar to already stored items, will activate the shared features and inherit the already existing learned patterns, ultimately leading to some generalizations (Pinker, 1998: 8). Since memory has connectionist properties, similar to those of the Rumelhart & McClelland's (1986) model, it is capable of capturing the semi-regular patterns among irregular verbs. Such properties would account for findings indicating that people have the ability to create irregular-sounding novel words (Pinker, 1998: 6-9).

Assuming a dual organization of the language faculty, the U-shaped developmental curve in first language acquisition can be explained in the following way. To begin with, children have no prior knowledge of the rules of a specific language. Lexical items are memorized as a lexical entry via exposure, including irregular items. Children begin producing frequent irregular forms correctly quite consistently. Through further input, they extrapolate the regular past tense rule, and thus begin marking the regular

past tense inflection. Co-occurring with the acquisition of the rule, overregularization errors start taking place. Words that have weaker memory traces, due to a lower frequency of exposure are more prone to overregularization. In the absence of a strong memory trace or exposure to a word, the regular past tense rule acts as a default, resulting in the overapplication of the rule. With time, and as the memory traces become stronger, these errors even out. When producing a past tense verb, the user first checks the mental lexicon to see if the entry only contains the stem, or if there is additional information stored, such as in the example (2) above. For irregular verbs, the productive rule is blocked when the irregular form is found in the lexical entry, resulting in the successful production of the irregular form. In cases of regular verbs, on the other hand, there are no irregular entries found which can block the rule, and the successful production of a regularly inflected word can thereby take place.

The extent to which a dual processing account is compatible with data from L1 experimental research is open to debate. Thus, Baayen & colleagues (2002) show that inflected regular verbs are both retrieved from the lexicon and computed on-line in parallel. At the same time, the study by Say & Clahsen (2002) provides support for the dual processing account by demonstrating that overregularized irregular verbs elicit stronger early negativity effects than wrongly inflected regular verbs.

1.3. U-shaped behaviour in SLA

Extant research on the plausibility of the dual processing account in L2 suggests frequency effects for irregular items, as reflected in faster and more accurate processing and acquisition, thus supporting a dual processing mechanism (Prasada, Pinker & Snyder, 1990; Seidenberg & Bruck, 1990; but see also critical discussion in Ellis, 2002). Further evidence of the plausibility of the dual processing account in the context of the L2 is provided by the study by Pliatsikas & Marinis (2013). They found a clear distinction between processing regular and irregular past tense forms of English verbs in a group of highly proficient L2 speakers. The authors attribute this distinction to the morphological difference between the two verb types and account for the documented delay in regular verb processing in terms of the activation of the regular rule, which is expected to trigger decomposition of the inflected form. Since irregular past tense forms are stored in the mental lexicon as separate entries, no computational processes are required, thus explaining the apparent facilitation in their RTs in that study.

Of specific interest for the current study is the extent to which L2 learners perform similarly on inflection tasks involving pseudo-verbs with similarity to regular and irregular verbs (Prasada & Pinker, 1993). Murphy (2004) used the design from Prasada & Pinker (1993) to compare L2 learners of English with different L1 backgrounds to

adult native speakers and a group of native English-speaking children (mean age = 8,9 years). The two adult groups performed similarly and replicated the original results obtained in the study by Prasada & Pinker (1993), supporting the model of dual mechanism in verb inflection. However, the child group differed significantly from the adult groups. Child participants inflected fewer items and displayed a similarity effect in both items patterned on regular and irregular verbs, reflecting an otherwise attested difference between children and adults in learning style (Johnson & Newport, 1989). Even though the adult results suggest that a dual mechanism might be common in both L1 and L2, Shirai (2019) is sceptical and observes that similar behavioural results (e.g., accuracy) do not necessarily reflect the same underlying cognitive mechanism in both populations.

The acquisition of morphology has been addressed extensively in the domain of second language, acknowledging its crucial role in L2 proficiency and why it may present problems for the L2 learner (DeKeyser, 2005, among others). According to the Bottle Neck hypothesis (Slabakova, 2008; 2013), morphology is one of the most challenging domains in the process of acquiring a second language. While U-shaped trajectories in acquiring a second language have been stipulated in domains where two competing forms render the same content (e.g., preterite forms of the verb, negation and *do-* support, McLaughlin, 1990), experimental research on U-shaped learning in SLA is still limited. Studies which report changes in accuracy trajectories diverge in both methods of data collection and target phenomena, and often report controversial evidence, which makes between study comparison difficult. Furthermore, a variety of terms have been used to define what might be otherwise defined as U-shaped behaviour in SLA ranging from non-linearity, restructuring, to variation and destabilisation. In addition, the reported evidence of what might look like U-shaped curves in learning a second language may have been caused by external factors, such as erroneous forms in the input provided (e.g., Wode, 1978).

Kellerman (1985) was among the first to identify changes in second language learners' production behaviour as U-shaped behaviour, similar to what has been reported in first language acquisition. His study addressed verb alternation patterns in Dutch learners of English of different ages. U-shaped behaviour was observed in the participants' acceptability judgements of transitive-intransitive alternations in English. The adolescent learner group rejected correct target language structures, while both younger and older participants accepted both transitive and intransitive uses. A possible account for this behaviour can be sought in the relatively lower frequency of the intransitive construction in English. With more exposure and resulting restructuring of the L2 knowledge, the more advanced learner group may no longer be influenced by L1 transfer.

The study by Shirai (1990) focused on lexical acquisition and investigated U-shaped learning among Japanese EFL learners of English of three different proficiency groups, along with one group of American native speakers of English. Shirai (1990) claimed U-shaped behaviour in L2 lexical development differs from U-shaped trajectories in L1 or other U-shaped curves in SLA, as it is driven by a growing L1-independence. No clear pattern emerged from the data in that study. Regarding lexical development, Shirai (1990) reported results indicative of U-shapes in certain contexts. Furthermore, he stated that documenting U-shaped behaviour in L2 lexical acquisition is highly item-specific and unpredictable (Shirai, 1990: 6). This author also proposed a three-stage model, where a transitional period from L1-dependency to L1-independency is intermediated by a restructuring stage, causing a U-shaped behaviour in lexical development.

Restructuring of the language system has been proposed to underlie language acquisition specifically in bilingual learners (Grosjean, 1992) and is often used in SLA research as well. Gass & Selinker (2008) hypothesize that restructuring of knowledge is the main cause of the three stages of U-shaped behaviour and learning. Knowledge restructuring can account for L2 morpheme production data, paralleling the three-phase model originally proposed by Karmiloff-Smith (1986) for child first language acquisition. The initial stages are input-driven, while later stages rely on representations and top-down processes. Further support for this idea comes from findings in other domains, such as article use in English as an L2 (Lu, 2001), *-ing* morphology (Lightbown, 1983), plural inflection for nouns (Wode, 1978).

Closer to the purposes of our study, Lightbown (1983) discussed data from French learners of English in a classroom context. She examined the use of the *-ing* progressive form in English at different proficiency levels (i.e., sixth, seventh, and eighth grade learners) and hypothesized that initially these learners were presented only with the progressive form. In stage one, with no other comparable verb form in their system, L2 learners over-extended the use of the progressive into contexts in which the simple present would have been appropriate. In stage two, when the simple present was introduced, learners not only had to learn this new form, but also readjust their information about the present progressive, and its limits. Because of the confusion and subsequent readjustment and restructuring of the progressive, there was a decline in both use and accuracy. Overtime, learners eventually restructured their L2 knowledge appropriately and were able to use both the progressive and the simple present in target-like ways.

More recently, the study by Geeslin & Guijarro-Fuentes (2006) addressed the acquisition of the Spanish copula contrast *ser* and *estar* (both equivalent to English "to be", however contrasting aspectually) among a native Spanish-speaking group (N=19)

and a native English-speaking group of second language learners of Spanish (N=7). The age of the participants ranged from 20 to 46 for the native Spanish-speaking group, and 20 to 47 for the native English-speaking group. The two groups were presented with multiple choice contextualized preference tasks. The Spanish-group was only tested once while the English-group was tested on four occasions during a 3-year degree program in Spanish. Using the Spanish-group's responses as a standard for comparison, the responses from the English-group were scored accordingly. Geeslin & Guijarro-Fuentes (2006) found patterns of U-shaped learning in the development of the copula choice for the English-group of second language learners of Spanish. Interestingly, a sharper U-shape curve was evident for a sub-group of the native English-speaking participants who studied abroad in Spain for 4 months.

Kounatidis (2016) reports results from a longitudinal study of intermediate proficiency level students of English. The design elicited the production of present tense forms of copular *be*, preterite forms of verbs and negation tested at 3 time points. The results indicated a U-shaped curve for copular *be*, production of past tense forms of familiar regular verbs, and an increasing slope for both familiar and novel irregular forms over the period studied. While accuracy for the familiar regular forms was close to ceiling at all 3 stages, the production of novel regular verbs declined consistently. This latter result, however, combined with the increase in accuracy rate in irregulars, is not consistent with the classical evidence in support of U-shaped trajectories in past tense acquisition by children in the L1.

Based on a large dataset, Casani (2020a; 2020b) provides relevant evidence of the cross-linguistic universality of U-shaped trends in the acquisition of morphology. In a detailed survey of grammar errors in a learner corpus of Italian as L2, Casani documents an inverted U-shape specifically in the acquisition of verbal (inflections) and nominal morphology (agreement, the article) across four levels of proficiency.

One of the biggest differences between first and second language acquisition is the existence of knowledge of an additional language (the L1) in an L2 learning context. During first language acquisition, one does not have any prior knowledge of any specific language, whereas the opposite is true for L2 acquisition. If we observe a curvature in accuracy and overregularization rates similar to the classic U-shaped curve, this suggests the presence of a more generic underlying learning pattern also in L2.

1.4. The current study

On the backdrop of research presented above, we ask the following research questions:

1. Do Norwegian L2 learners of English exhibit a similar learning curve in relation to the acquisition of irregular verb morphology as (child) L1 learners of English?
2. Can this evidence suggest a common learning pattern for L1 and L2 acquisition?

We expected to find significant differences in accuracy and overregularization rates between three groups of L2 learners at different levels of proficiency. The *U* in the term *U-shaped development* models the dip in accuracy. However, this learning pattern involves two related processes: a dip in accuracy followed by subsequent rise in accuracy accompanied by an increase in rate of overregularization and a subsequent decrease of overregularization. Thus, U-shaped development can be documented by following either the accuracy curve or the overregularization one. Although the latter represents an inverted U-shape, it is an inherent part of the phenomenon referred to as U-shaped development. Observing the entire U-shaped curve in the current study would therefore involve an increase in the rate of overregularization errors for the middle proficiency group, in comparison to the lower and higher proficiency groups. In addition, the overall curvature shape can be broken down into smaller segments reflecting different stages of the process, like observed in L1 grammar inflection learning. Furthermore, accuracy rates and overregularization rates are linked together, and any results should reflect both aspects simultaneously.

2. Methodology

2.1. Participants

The participants in this study were students at a lower-secondary school in Norway. They were recruited from three different grades: 8th grade (N=20), 9th grade (N=24), and 10th grade (N=23), respectively. English skills in Norwegian children and adolescents follow a similar trajectory, and seem to accelerate around lower secondary school (Dahl & Vulchanova, 2014). Therefore, we used school grade as a proxy for English proficiency level. However, in line with the Norwegian school system in general, and, more importantly, in accordance with the European Framework for Languages (https://www.coe.int/en/web/portfolio/home), the learner groups were further classified as follows: 20 participants in grade 8 (B1 level of general proficiency in English), 24 participants in grade 9 (B2 level) and 23 participants in grade 10 (C1 level).

Because irregular past tense morphology is learned by rote and using memorization sheets at school (Dahl, 2014), and vocabulary size increases steadily, but slowly in

Norwegian school children (Sivertzen, 2013), we estimated that grades 8-10 would be optimal to test for U-shaped learning curves. Ages 10-15 years mark an increase in metalinguistic awareness, as documented by Evensen (2014). Thus, an additional rationale for using these age groups was the development of L2 metalinguistic skills, which may be necessary for performance on the task. To confirm these assumptions, we interviewed the school-teachers of the participating groups. Several of the teachers reported a decline in overregularization errors during the last year of lower secondary school, consistent with our participant selection. In Norway, grade 8-10 attend the same school. Prior to this, learners receive mostly explicit instruction and memorize irregular past tense verbs through memorization sheets. Later through grades 8-10, however, most instruction relating to irregular verbs is in the form of corrective feedback and exposure to target language texts. L2 learning becomes increasingly driven by exposure to authentic input both at school and outside of school via extensive media use (Brevik & Rindal, 2020).

Participants with known learning deficits or other relevant developmental deficits, such as dyslexia, were excluded. Other exclusionary criteria were balanced bilingualism, having a different L1 than Norwegian, and poor performance on the experimental task (i.e., below 48% accuracy for overall verb production). The final participant count was as follows: 8th grade at B1 level (N=17), 9th grade at B2 level (N=19), and 10th grade at C1 level (N=15).

Participants signed a written informed consent to participate in the study. For participants in the 8th and 9th grade who were younger than 16, consent had to be given by their parents. A questionnaire related to the participant's language background, and potential presence of learning deficits, was also administered to parents and participants, respectively.

Students in Norway begin learning English formally during their first year of attending school (at age 6 years old). After completing the 7th grade and upon entering the lower-secondary level, they will have received 366 hours of education in English, whereas during their time at lower-secondary school, they will receive an additional 228 hours (Utdanningsdirektoratet, 2013). Normally, students in the 8th grade are between 13 and 14 years old, 14 and 15 years old in the 9th grade, and 15 and 16 years old in the 10th grade.

After the research project was approved by both the school and the Norwegian Centre for Research Data (henceforth NSD), a select group of volunteer teachers helped grant access to their classes.

2.2. Stimuli

Participants were given two types of elicitation tasks modelled after Berko's (1958) Wug-test. One pertained to verb inflection knowledge, whereas the other one was related to nominal inflection knowledge (not reported here). To create the items for the verb inflection knowledge task, we began by making an overview of the learning material used by the school for teaching English. We identified three books which the school used primarily for each grade. First, we compiled a list of the irregular verbs found across all books' glossaries. In total, 93 irregular verbs were found. Examples such as "to set", "to come", and "to wake" illustrate a few of the irregular verbs that we identified.

Two conditions based on frequency were created, as previous studies indicate frequency of exposure to be an important factor in relation to U-shaped learning and the production of overregularization errors (Marcus et al., 1992). The two conditions were labelled "easy condition" corresponding to high frequency items, and "hard condition" corresponding to low frequency items, respectively. We checked the frequency ratings for all the verbs in BYU's *iWeb: The intelligent Web-based Corpus*. Because children and adolescents are frequently exposed to the internet (Vulchanova et al., 2017), we chose this web-based corpus based on the assumption that the verb frequencies in the corpus will approximate the frequencies they are exposed to in their daily lives. An open search for verbs, grouped by lemmas, resulted in a list of the top 1000 most frequently occurring verbs in the corpus. Of the 93 irregular verbs found across the books, 20 with a frequency ranking smaller than 50 out of the 1000 hits (RAW FREQ greater than 6056787) were chosen as items for the easy condition (e.g., "to run", "to know", and "to take"). Next, 20 of the irregular verbs with a frequency ranking greater than 300 of the 1000 hits (RAW FREQ smaller than 1013581) were chosen as items for the hard condition (e.g., "to sleep", "to freeze", and "to bite"). The selection process resulted in an elicitation task containing a total of 40 items with irregular inflection. Item presentation followed the same order for all participants. Out of the 40 verbs, 17 were cognate Germanic verbs (e.g., swim, come), and 12 out of these were strong verbs in both the L1 and L2, with 7 out of the 12 inflecting in a similar way. Thus, 7 items out of 40 were cognate, strong and inflected similarly for the preterite (e.g., sing*/synge*; sang*/sang*). Observe, however, that in most of these cases these forms are not equivalent phonologically.

The following example in (3) illustrates a stimulus set up:

(3) STEAL: Paul likes to steal things.

However, during the party, Paul _____ Lisa's heart.

The target word was specified in its non-finite form for the verbs (spelled out in capital letters to highlight that this is the target word) and was also used in the present tense in the introductory sentence. In the second sentence, a blank was left where the participant was asked to fill in the irregular past tense form of the word according to context, which in the above example would have been "stole". The Wug-test has been documented to reliably elicit target item production in both children (Berko, 1958) and adults (Vulchanova et al., 2012) in various experimental set ups.

2.3. Procedure

Prior to any testing, and before the participants received any information, consent sheets were given to both parents and participating students. Ahead of each testing session, a list of participants was created from the consent sheets collected by the teachers involved.

Two separate online response forms were created to accommodate participants who either needed parental consent or who were able to give consent themselves using *SurveyGizmo* as platform. We also collected reaction times to each question. These measurements provided information on the amount of time each participant spent on answering each item of the elicitation task.

Each testing session took about 60 minutes and was held at the school during teaching hours. Some sessions took place in a mixed classroom environment with other non-participating students present, while others were held in a room separate from the non-participating students. During mixed classroom sessions, non-participants were given individual work by the teacher and encouraged to keep quiet, so as to not disturb the participating students. All testing sessions were administered by the first author of this paper.

2.4. Analysis

For the analysis, grade was used as the independent variable, whereas accuracy and overregularization were used as dependent variables. All responses to items in the elicitation task were first compiled according to three categories: accurate responses, overregularization errors, and other types of errors. Percentages were calculated for each variable. For each variable, the mean, standard deviation, and 95% confidence interval were calculated for each group/grade.

The results focus on the relation between grade and rate of accuracy, as well as grade and rate of overregularization errors. A Shapiro-Wilk test revealed that the data

did not follow a normal distribution. Consequently, the Wilcoxon test was adopted, as it does not assume normality. This test was chosen to assess the degree of overlap, or lack thereof, between the distributions to see if they were significantly different or not.

We additionally performed a Pearson-correlation test between the participant's self-rated level of proficiency and accuracy, in addition to overregularization rates, to see whether these variables were related or not.

3. Results

3.1. Results by group and category

Table 1 summarizes the descriptive statistics for each group. A Shapiro-Wilk test for normality was also applied to each variable. Although some of the variables did distribute normally, most did not. Consequently, interpreting the confidence intervals must be met with caution.

Table 1: Descriptive statistics from analysis based on grade

		Mean	*SD*	*Confidence interval 95%*		*Shapiro-Wilk test*
				Min	*Max*	
	Grade:					
Overall accuracy						
	8th	75.0	12.0	68.8	81.2	$p = 0.786$
	9th	77.2	11.9	71.5	83.9	$p = 0.276$
	10th	86.8	8.4	82.2	91.4	$p = 0.01$
Accuracy, easy condition						
	8th	86.5	7.9	82.4	90.5	$p = 0.702$
	9th	88.4	10.9	83.2	93.7	$p = 0001$
	10th	93.0	6.2	89.6	96.4	$p = 0.003$
Accuracy, hard condition						
	8th	62.9	17.9	53.8	72.1	$p = 0.411$
	9th	65.5	15.8	57.9	73.1	$p = 0.393$
	10th	80.3	12.7	73.3	87.4	$p = 0.096$
Overall overregularization						
	8th	7.7	8.1	3.5	11.9	$p = 0.001$
	9th	9.7	7.1	6.2	13.1	$p = 0.009$
	10th	5.8	4.5	3.3	8.3	$p = 0.006$
Overregularization, easy cond.						
	8th	2.9	3.1	1.4	4.5	$p = 0.0004$
	9th	3.2	4.2	1.2	5.2	$p = 8.1133e-05$
	10th	1.0	2.1	-0.1	2.1	$p = 3.481e-06$
Overregularization, hard cond.						
	8th	12.1	14.3	4.7	19.4	$p = 0.001$
	9th	15.8	12.4	9.8	21.8	$p = 0.061$
	10th	10.0	8.2	5.4	14.6	$p = 0.005$

Note: Numbers are presented in percentages converted from fractions.

The mean for overall accuracy rate between the three grades increases from 75% to 96.8%. In contrast, the rate of overregularization increases between grade 8 and 9, and then drops from 9th to 10th grade.

As seen below, Figure 1 and Figure 2 provide a visualization of the data in boxplot format, while Figures 3 and 4 provide a fitted line plot for participants' accuracy and overregularization rates, necessary to assess the learning curves in our data. A classical U-shaped scenario would see a decline in performance in one domain (i.e., irregular past tense morphology), despite overall increase in general proficiency of the participants. Thus, one would first observe error-free performance, followed by performance with errors, which ends up with error-free performance again, consistent with overall assumptions in the literature (e.g., Kellerman, 1985). Our data, however, show similar accuracy rates for the two lower-level proficiency groups, and an increase in accuracy for the highest level proficiency group (Figure 3). As seen in Figure 1 and Figure 2, the 10th grade group performed at ceiling for frequent irregular verbs (easy condition), but showed variation for the less frequent category (hard condition). However, the overregularization errors display an inverted U-shaped trajectory (Figure 4).

Figure 1. Accuracy scores for easy (turquoise) vs hard (red) condition for verbs, by grade

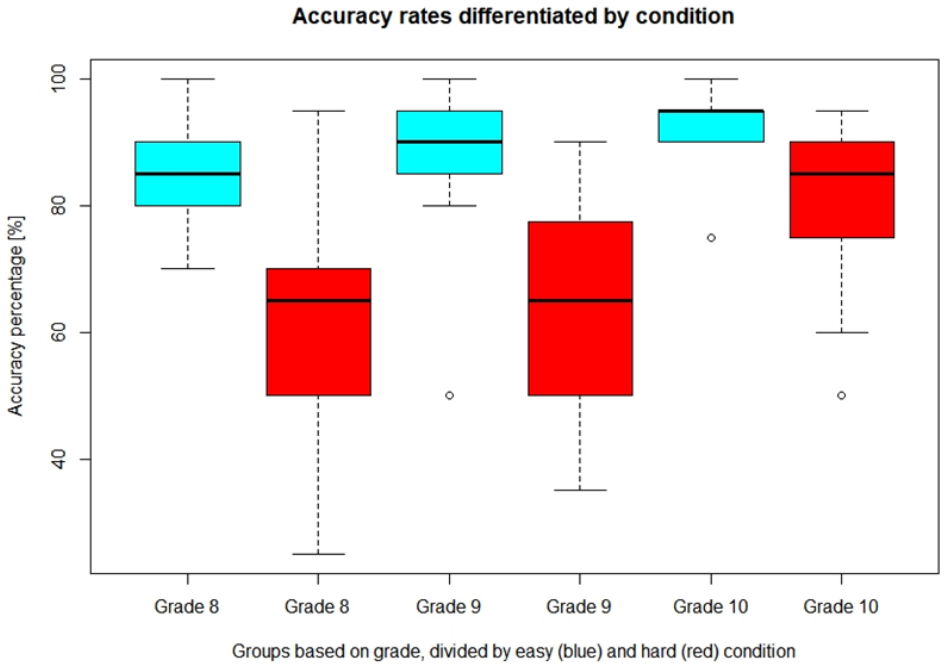


Figure 2. Comparison between easy condition (turquoise) and hard condition (red) for rate of overregularization of verbs, by grade

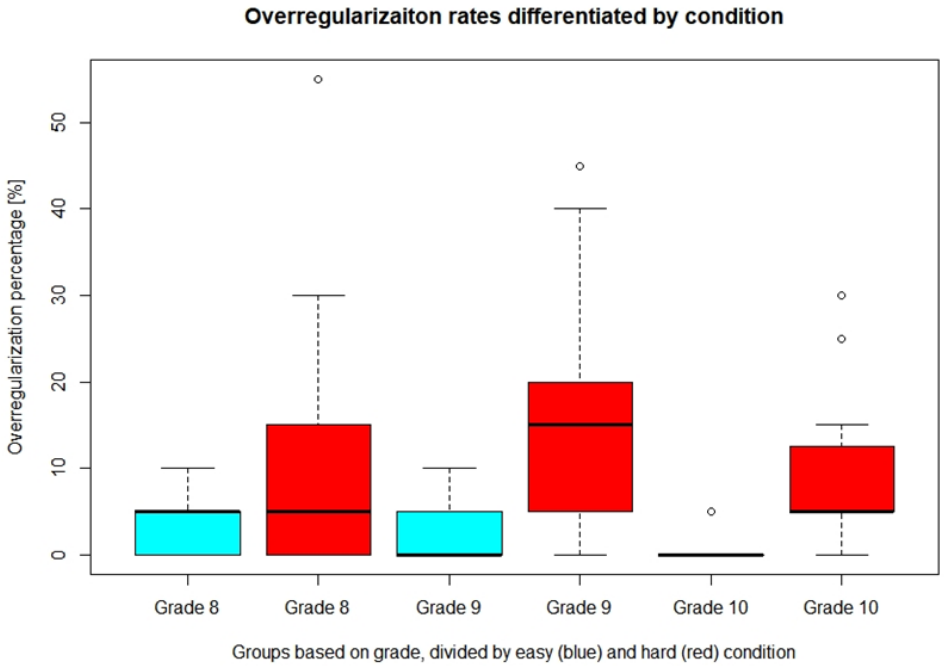


Figure 3. Fitted line plot for overall accuracy scores between grades

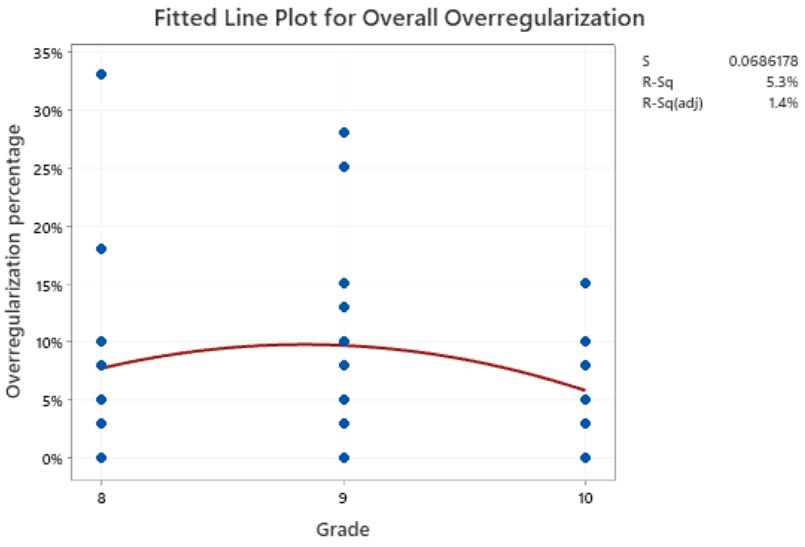
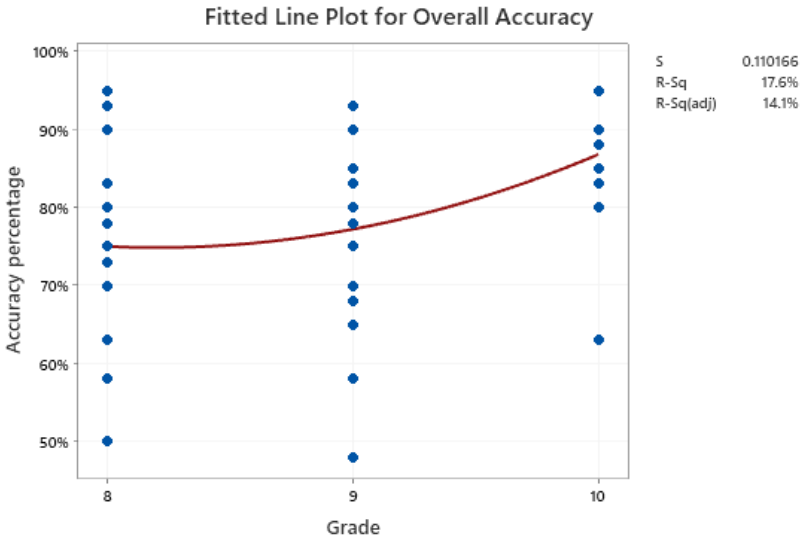


Figure 4. Fitted line plot for overall overregularization rate between grades



There were significant differences between the two conditions within two of the groups (8th and 10th grade), but not for the middle group (9th grade). This was confirmed by a Wilcoxon test which gives an indication of whether two distributions (X and Y) overlap or not, and if they differ significantly, thus testing the null hypothesis for each group compared to each other group. For exact *p*-values, see Table 2. A post hoc Bonferroni correction (= 0.016) for multiple comparisons confirmed the significance effect.

Table 2. Within-group differences between easy and hard condition

Variable tested:	Grade:	*W*:	*p-value*:
Easy & Hard condition			
	8th grade	255	0.0001386
	9th grade	326.5	1.849e-05
	10th grade	190.5	0.001033

Accuracy on the frequent group of verbs (easy condition) lies above 85% for all participant groups. The less frequent verbs (hard condition) elicit poorer responses at 62% (8th grade) and 65% (9th grade), with a steep rise to 80% only for grade 10. In

addition, there is much less variation in the easy than the hard condition within each group, as indicated by the interquartile range, the median and the tails in the boxplots. Performance appears to increase in the hard condition across the three groups, but mostly in-between grades 9 and 10. The same is true of the variation in accuracy, with greater variation displayed by the two lower proficiency groups (8th and 9th grade), and much less variation in the highest proficiency group (10th grade). Although the rate of overregularization errors seems stable between 8th and 9th grade within the easy condition of verbs, the medians differ, and this is confirmed by the results from the Shapiro-Wilk test (Figure 2). Furthermore, the 10th grade group exhibits a floor effect for this kind of error for highly frequent verbs. Figure 2 also indicates that the rate of overregularization within the easy condition for the 10th grade is tightly distributed (with almost no variation). Altogether, every grade produced more overregularization errors for less frequent verbs (hard condition) in comparison to frequent ones (easy condition). We also see that the rate of overregularization errors in the hard condition rises in the 9th grade to drop again in the 10th grade, thus forming an inverted U-shape, evident in both Figure 2, and in the curve plot in Figure 4. The fitted line plots in Figure 3 and Figure 4 display opposite trends: an increasing slope in Figure 3 (rise in accuracy) and an inverted U-shape in Figure 4 (drop in overregularization).

Since some of the data did not distribute normally, the non-parametric unpaired version of the Wilcoxon test, which does not assume a normal distribution, was applied in R to see whether the distributions overlapped or not. Table 3 provides the details with group comparisons.

Table 3. Wilcoxon tests

Variable tested:	Grades being compared:	*W*:	*p-value*:
Overall accuracy			
	8th & 9th grade	137	0.445
	9th & 10th grade	66	0.008
	8th & 10th grade	48	0.002
Accuracy, easy			
	8th & 9th grade	123.5	0.224
	9th & 10th grade	98	0.113
	8th & 10th grade	62	0.011
Accuracy, hard			
	8th & 9th grade	144.5	0.597
	9th & 10th grade	65	0.007
	8th & 10th grade	53	0.004
Overall Overregularization			
	8th & 9th grade	127	0.271
	9th & 10th grade	195	0.066
	8th & 10th grade	141	0.614
Overregularization, easy			
	8th & 9th grade	165	0.916
	9th & 10th grade	180	0.119
	8th & 10th grade	171	0.054
Overregularization, hard			
	8th & 9th grade	120	0.187
	9th & 10th grade	181.5	0.172
	8th & 10th grade	124	0.907

After running a post hoc Bonferroni correction for multiple comparisons (α = 0.0083) for overall accuracy as well as accuracy in the easy and hard condition, the distributions for 8th and 9th grade did not differ significantly. The distributions, when comparing 9th and 10th grade for accuracy in the easy condition, did not differ either. However, significant differences between the 9th and 10th grade were found in both the

overall accuracy in verb production as well as the accuracy rates in the hard condition. These differences were also observed between the 8th and 10th grade performances. For the accuracy on the easy condition, the only significant difference was between the 8th and 10th grade. In sum, the 8th and 9th grade do not differ in accuracy. The 9th and 10th grade differ on overall accuracy and on less frequent (hard) verbs, whereas the 8th and 10th grade differ systematically. Taken together, these results indicate an increase in accuracy, primarily between the two younger (lower proficiency) groups and the older, highest proficiency group.

As for overregularization errors, the distributions overlap and did not differ significantly for the 8th and 9th grade ($p = 0.271$) on overall overregularization rates, and for the 8th and 10th grade ($p = 0.614$), which is to be expected from a U-shaped learning pattern. A trend to significance in overall overregularization errors can be observed when comparing the 9th and 10th grade ($p = 0.066$). A similar trend to significance can be observed for overregularization rates in the easy condition, for the 8th and 10th grade ($p = 0.054$). Note that none of the tests could yield exact *p*-values, due to ties in the data, which means that all the *p*-values are approximate computations.

Reaction times were recorded for all responses. Table 4 provides the descriptive statistics for the reaction times on accurate and overregularized responses, whereas Figure 5 and Figure 6 visualize the data.

Table 4. Descriptive statistics for reaction times

		Mean	*SD*	*Confidence interval 95%*	
				Min	*Max*
	Grade:				
Reaction times for accurate Responses	8th	20.6	7.1	16.9	24.2
	9th	16.2	7.7	12.5	19.9
	10th	13.0	4.2	10.6	15.3
Reaction times for overregularized responses	8th	33.2	25.0	20.3	46.0
	9th	18.7	10.4	13.7	23.7
	10th	15.7	7.1	11.8	19.6

Note: Time is presented in seconds.

Figure 5. Average time spent on accurate answers with standard deviation error bars. Average time spent on accurate verb responses

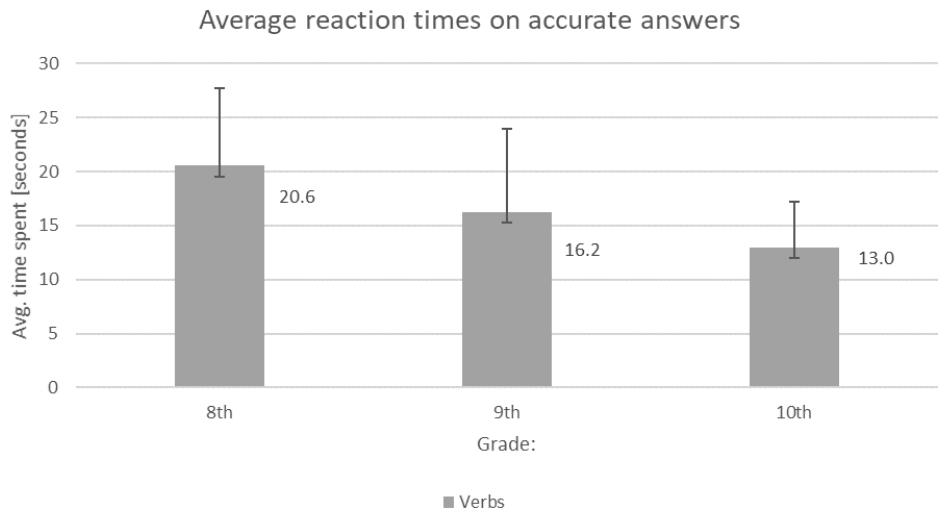
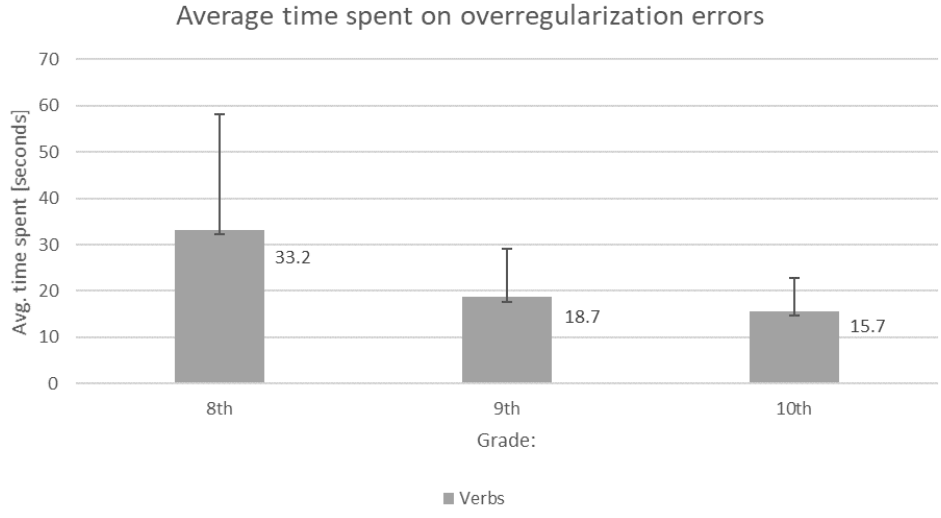


Figure 6. Average time spent on overregularization errors, including error bars for standard deviation. Average time on overregularization errors for verbs



As Figure 5 illustrates, reaction times on accurate responses decrease with grade, suggesting an increase in proficiency and faster lexical retrieval of correctly produced verbs. A similar trend is observable for overregularization errors (Figure 6). This

latter trend is relevant for the current hypothesis. The decrease in time spent on overregularization suggests a process of restructuring the target language knowledge as a result of acquiring the rule. As shown by the standard deviation error bars, the variability also seems to decrease with grade for time spent on verb production in both figures.

Table 5. Pearson-correlation between participant's self-rated level of proficiency and rates of accuracy and overregularization

Variable:	Grade:	*Correlation:*	*p-value:*
Overall accuracy			
	8th	0.2761288	0.283
	9th	0.3016911	0.209
	10th	0.3591946	0.188
Accuracy, easy cond.			
	8th	0.4424952	0.075
	9th	0.3012573	0.210
	10th	0.372678	0.171
Accuracy, hard cond.			
	8th	0.1908335	0.463
	9th	0.2343126	0.334
	10th	0.3027247	0.272
Overall overregularization			
	8th	0.001714255	0.994
	9th	-0.2487454	0.304
	10th	-0.4861471	0.066
Overregularization, easy cond.			
	8th	-0.2362087	0.361
	9th	-0.3860624	0.102
	10th	-0.559017	0.030
Overregularization, hard cond.			
	8th	0.05121768	0.845
	9th	-0.1553049	0.525
	10th	-0.4214636	0.117

An additional Pearson correlation test was performed to see whether the participants' self-rated level of proficiency correlated with accuracy rates, as well as the rate of overregularization errors. A Bonferroni was calculated ($\alpha = 0.0083$) to adjust for multiple comparisons. No significant correlations were found between the participants' self-rated level of proficiency and accuracy rates. As for the overregularization errors, no correlation was found between the participants' self-rated level of proficiency and the overall rate of overregularization errors of verbs (shown in Table 5). However, we did find a significant ($p = 0.03028$) negative correlation ($r = -0.559017$) between the participant's self-rated level of proficiency and rate of overregularization within the easy condition.

In the next section we report the individual results for each of the groups.

3.2. Individual results

Table 6 summarizes individual performance and number of participants at or below chance level (.50), and number who performs above (for both conditions). Further, the table provides details the number of participants who performed above and under the group's average performance.

Table 6. Individual results

		8th grade		*9th grade*		*10th grade*	
		Easy cond.	Hard cond.	Easy cond.	Hard cond.	Easy cond.	Hard cond.
Chance level conditions	50% both						
	Above:	17	12	18	13	15	14
	At chance:	0	4	1	4	0	1
	Under:	0	1	0	2	0	0
	Total:	17	17	19	19	15	15
No. of participants Under/above mean							
	Above mean:	8	10	13	8	9	8
	Under mean:	9	7	6	11	6	7
	Total:	17	17	19	19	15	15

The results in Table 6 provide evidence of overall good individual performance in all groups with very few participants performing at or below chance (observed only in the hard condition). We also see an increase in performance accuracy, with fewer individuals performing below the mean in both conditions with grade. These results thus complement the individual variation picture presented in the box-plot analyses above.

4. Discussion

Prior to testing, we expected to find significant differences in the levels of accuracy and overregularization errors when comparing the three learner groups in this study, in addition to possibly finding ceiling effects in the most proficient group (i.e., 10th grade at C1 level). We assumed evidence for U-shaped learning would support an underlying input-driven learning pattern shared between L1 and L2 acquisition. We hypothesized a U-shaped trend would be evidenced by an uneven (i.e., curved) rate in performance, with a dip in accuracy and increase in overgeneralization rates in the middle proficiency group, as opposed to a linear increase/decrease trend. With regards to accuracy, we observed a clear increase in performance between the 8th and

the 10th grade, suggesting that knowledge of irregular inflections increases with time, and between the least (B1) proficient and the most (C1) proficient group. Our results are thus consistent with the data in Kounatidis (2016) on past tense forms of irregular verbs.

A significant point of interest in this study was the trend observed at the mid-point (9th grade at B2 level). Indeed, a difference in overall accuracy was observed between the 9th grade and the 10th grade, driven primarily by performance on less frequent (hard) verbs. Concerning overregularization, we observed trends toward significance. In particular, this was found in overall performance between the 9th and the 10th grade, as well as between the 8th and the 10th grade on the frequent (easy) verb condition. These differences suggest the 9th grade was a mid-point in the study (≈ middle proficiency), with a clear difference between that point (level) and the most advanced one. We interpret these findings to be indicative of part of the U-shaped curve, modelling the later stages of the phenomenon, thus providing a parallel to behaviour observed in L1 acquisition (Marcus et al., 1992).

4.1. Evidence for U-shaped learning in the L2

The first question in this study sought to determine whether Norwegian L2 learners of English exhibit U-shaped learning similar to L1 learners of English when acquiring irregular aspects of verb morphology in English. We interpret our results as suggestive evidence in favour of a U-shaped trend. This interpretation is limited to a small interval from the later stages of U-shaped learning during L2 acquisition among Norwegian learners of English. During the later stages of U-shaped learning, the dual-mechanism hypothesis predicts that during U-shaped development, the level of accuracy should steadily increase, while the rate of overregularization errors decrease. Our findings model this trend.

Firstly, overall accuracy rates improve with each grade, rising from 75% correct responses to 86.8% on average. This is confirmed by the systematic difference in accuracy between the 8th and the 10th grade (see Table 1). In addition, the variation in accuracy decreases with grade and proficiency level. This decrease in variability, in conjunction with the increase in performance accuracy, is indicative of an overall increase in proficiency among participants across grades. However, performance does not seem to increase in a linear fashion with grade. Instead, the middle point (i.e., the 9th grade), evidences a rise in overregularization errors combined with no significant difference between accuracy on frequent (easy) verbs and less frequent (hard) verbs (Table 2). This suggests an input-driven restructuring of grammar competence in this group, as reflected in the trend to significance on accuracy on less frequent verbs and

overall overregularization errors between the 9th grade and the 10th grade (Table 3). Importantly, our findings not only document that L2 learners overregularize in the process of the acquisition of L2 morphology, but also demonstrate that this happens at a specific time-point in the course of learning. This shows the process is not linear, thus suggesting an inverted U-shape curve.

These results seem to better reflect the participants' development regarding their proficiency level than their self-reported level of proficiency. A Pearson correlation-test between participants' self-rated level of proficiency and the rate of overregularization errors did not reveal any relationship (Table 5). If anything, the correlation between these two factors seemed to decrease with grade, replicating well-known trivial effects between level of performance and self-assessment (Kruger & Dunning, 1999; but see LeBranc & Painchaud, 1985 for evidence to the contrary). Furthermore, participants' self-rated level of proficiency did not seem to be a reliable measurement in our study. When asked to rate their overall level of proficiency on a Likert-scale ranging from 1-5, participants did not use the whole range, but instead only reported values between 3-5, compromising the correlation analysis results.

Secondly, participants were sensitive to frequency effects. There was a clear frequency effect observed in performance on highly frequent items (easy condition) and the low frequency ones (hard condition), consistent with the results in Murphy (2004). Not only was this effect visible for accuracy (Figure 1), but also for rates of overregularization errors (Figure 2). This observation supports Pinker's (1998) idea of the links between frequency of exposure and memory traces and their role in the lexical retrieval of words. A higher frequency of exposure leads to stronger memory traces. This means that they are less prone to overregularization errors and more likely to be correctly retrieved from the mental lexicon as a first step in the process of producing the target inflection.

Thirdly, there were some observable curves between the grades in overregularization rates in overall verb production (Figure 4), and some interesting differences became evident when differentiating between conditions. In the easy condition, a floor effect was observed for the 10th grade (Figure 2) suggesting a dramatic drop in frequent verb overregularization in the most advanced participant group. In addition, when comparing median values between grades in the hard condition, we saw a rise and drop in overregularization rates resembling an inverted U-shape. This was not confirmed, however, by the Wilcoxon test results.

Finally, reaction times for accurate answers and overregularized items decreased with grade. Not only does the average time spent on these types of responses decrease with grade, but also the variability (see SD and mean values in Table 4). In addition, participants across all grade levels spent more time on overregularized responses than

on accurate responses. This is yet further evidence suggesting an overall increase in proficiency from grade to grade and is consistent with the findings in Pliatsikas & Marinis (2013) of increased reaction times for regular verb forms as a result of an underlying process of decomposition.

We have reason to believe that our data speak in favour of input-driven U-shaped learning. The first indication is connected to the frequency effects observed in all results. Although our participants are L2 learners of English, their responses parallel the frequency effects expected of the forms in the target language English. Thus, the results indicate sensitivity to the target language and to English item frequencies. It should be noted here that the stimuli verbs were neither cognates of Norwegian verbs, nor corresponded in paradigm regularity between the two languages (e.g., the English irregular may correspond to a Norwegian regular verb: "to make" – *å lage*, and "to think" – *å tenke*). This comes to suggest that the results reported here reflect acquiring English as a (target) language in the later stages of the U-shaped curve. As grade and proficiency increase, the variability in all measurements decreases and the participants begin to produce language more uniformly. This parallels the development among L1 speakers from adolescence to adulthood. Although the two languages are typologically similar in their division between a productive past tense rule and sub-regular patterns among irregular verbs, the items used in this study did not have a one-to-one correspondence on inflection paradigm type between the two languages, with only 7 items inflecting similarly in both languages. This similarity, however, did not prevent participants from over-regularizing, as seen in the non-target form "frozed". The current findings can be taken as support for an overall learning trend rather than cross-linguistic influence (Casani, 2020a, 2020b). Since Casani's (2020a, 2020b) studies also document a similar trend for basic syntax (word order and negation), as well as for prepositions and conjunctions, the question arises whether U-shaped trajectories can be more universal than originally thought and can be documented across the board for second language acquisition and across languages. Such evidence will necessarily also call for a revision of our theory of the nature of U-shaped trajectories in language acquisition.

5. Conclusion

This study set out to explore whether Norwegian L2 learners of English exhibit a similar learning curve in the acquisition of irregular and regular aspects of verb morphology as L1 learners of English do. Our evidence suggests that our group of Norwegian L2 learners of English do in fact display later stages of U-shaped learning behaviour in the acquisition of verbs. Accuracy rates increased with grade, whereas overregularization rates decreased. Furthermore, 10th grade exhibited a floor effect for the overregularization rate of highly frequent verbs. A Wilcoxon test found a trend to significance in rate of overregularization of highly frequent verb items between 8th

and 10th grade. This result can likely be ascribed to the small sample size, and we speculate that we could have found a clearer result and more power from a bigger sample. Our results speak in favour of the U-shaped trajectory as an underlying general pattern characteristic of language learning and related to the operation of storage (memory) mechanisms and the gradual acquisition of rules and pattern extraction to construction, as a result to exposure to input, and consistent with usage-based frameworks of language acquisition (Ellis, 2002).

Although we have suggestive evidence for an interval of a U-shaped curve, this interpretation must be met with caution. Our limited data (i.e., sample size, age groups) cannot confirm nor disprove the existence of classical U-shaped learning in the acquisition of L2 morphology. Consequently, this reduced our choice of statistical tests for our analysis. Another limiting aspect to our study is the amount of variability in many of the measurements and variables within each group. Separating and organizing the participants according to levels of proficiency may have yielded an even clearer picture of the developmental curvature that we are interested in. However, grouping participants according to proficiency level is not unproblematic either, since we lose the benefit of variability in the data. A final limitation is that we did not control for variation between British and American Englishes. Due to the diverse input which Norwegian learners of English receive and the resources we had, it was not feasible to control for this.

However, further studies are indeed necessary to confirm this. In particular, within-participant longitudinal designs, similar to Geeslin & Guijarro-Fuentes (2006), would be highly relevant to address this phenomenon in further detail. The current study should also be replicated including data from earlier grades at lower proficiency level to see if the entire U-shaped curve can be observed. In addition, testing English L2 learners, but with a typologically dissimilar L1, such as Mandarin Chinese, and analysing data based on L2 learner corpora from other second languages of the type reported in Casani (2020a, 2020b) would be important to establish the universality of the phenomenon. We also believe a study which compares Norwegian L2 learners with child Norwegian L1 speakers would provide important evidence of the nature of U-shaped trajectories and the underlying factors that impact on this pattern.

6. Acknowledgements

The authors would like to thank the participants and their teachers for their time and participation in the test. We would also like to thank Giosuè Baggio for his help with our statistics, Olav Landmark Pedersen for his expertise using R, and Brittany Arnold for her valuable feedback during our final revision.

7. References

Baayen, H., Schreuder, R., De Jong N., & Krott, A. (2002). Dutch Inflection: The Rules that Prove the Exception. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.). *Storage and Computation in the Language Faculty. Studies in Theoretical Psycholinguistics* (Vol 30, pp. 61-92). Dordrecht: Springer.

Berko, J. (1958). The Child's Learning of English Morphology. *WORD*, *14*(2-3), 150-177.

Bowerman, M. (1982). Starting to talk worse: Clues to language acquisition from children's late speech errors. In S. Strauss (Ed.). *U-shaped behavioral growth* (pp.101-145). New York: Academic Press.

Brevik, L. M., & Rindal, U. (2020). *Teaching English in Norwegian classrooms: From research to practice*. Oslo: Universitetsforlaget.

Carlucci, L., & Case, J. (2013). On the Necessity of U-Shaped Learning. *Topics in Cognitive Sciences*, *5*, 56-88.

Casani, E. (2020a). Valutare la competenza morfosintattica in italiano L2. Una validazione corpus-based dei livelli del QCER. In E. Nuzzo, E. Santoro, & I. Vedder (Eds.). *Valutazione e misurazione delle produzioni orali e scritte in italiano lingua seconda* (pp.15-26). Florence: Franco Cesati Editore.

Casani, E. (2020b). Corpus d'apprenants et recherche linguistique. Une étude sur l'italien. Actes du Colloque CEDIL'18 «Qu'a changé le numérique? Perspectives linguistiques et didactiques». Université Grenoble Alpes, 30 mai - 1 juin 2018. https://hal.univ-grenoble-alpes.fr/hal-02641560/document

Cazden, C. B. (1968). The Acquisition of Noun and Verb Inflections. *Child Development*, *39*(2), 433-448.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Dahl, A. (2014). *Young Second Language Learners: The acquisition of English in Norwegian first-grade classrooms*. PhD thesis. Department of Language and Literature, Norwegian University of Science and Technology.

Dahl, A., & Vulchanova, M. (2014). Naturalistic acquisition in an early language classroom. *Frontiers in Psychology*, *5*, 1-9.

DeKeyser, R. M. (2005). What Makes Learning Second-Language Grammar Difficult? A Review of Issues. *Language Learning*, *55*, 1-25.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2003). Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. In C. J. Doughty, & M. H. Long (Eds.). *The Handbook of Second Language Acquisition* (pp.33-68). Williston: Blackwell Publishing Ltd.

Ervin, S. M., & Miller, W. R. (1963). Language development. In H. W. Stevenson (Ed.) & J. Kagan, C. Spiker (Collaborators) & N. B. Henry, H. G. Richey (Eds.). *Child psychology: The sixty-second yearbook of the National Society for the Study of Education, Part 1* (pp. 108-143). Chicago: University of Chicago Press.

Evensen, M. W. (2014). *Searching for patterns in English as a second language: a Norwegian study of 10 and 15 year old learners*. Unpublished MA thesis. Department of Language and Literature, Norwegian University of Science and Technology.

Gass, S. M., & Selinker, L. (2008). *Second Language Acquisition. An introductory course* (3rd ed.). New York: Routledge.

Geeslin, K. L., & Guijarro-Fuentes, P. (2006). A Longitudinal Study of Copula Choice: Following Development in Variable Structures. In N. Sagarra, & A. J. Toribio (Eds.). *Selected Proceedings of the 9th Hispanic Linguistic Symposium* (pp. 144-156). Somerville, MA: Cascadilla Proceedings Project.

Grosjean, F. (1992). Another View of Bilingualism. In R. J. Harris (Ed.), *Advances in psychology* (Vol. 83). *Cognitive Processing in Bilinguals* (pp. 51-62). Elsevier Science Publishing Company Inc.

Jackendoff, R. (2002). *Foundations of Language. Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Johnson, J.S., & Newport, E.L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60-99.

Karmiloff-Smith, A. (1986). Stage/structure versus phase/process in modeling linguistic and cognitive development. In I. Levin (Ed.), *Stage and Structure: Reopening the Debate* (pp. 164- 190). Norwood, N.J.: Ablex.

Kellerman, E. (1985). If at first you do succeed... In S. M. Gass, & C. G. Madden (Eds.). *Input in Second Language Acquisition* (pp. 345-353). Rowley: Newbury.

Kounatidis, J. (2016). Destabilisation, IL variation and restructuring in foreign language learning. *Yearbook of the German Cognitive Linguistics Association*, 4(1), 261-288.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134.

LeBlanc, R., & Painchaud, G. (1985). Self-Assessment as a Second Language Placement Instrument. *TESOL Quarterly*, *19*(4), 673-687.

Lightbown, P. M. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. W. Seliger, & M. H. Long (Eds.). *Classroom Oriented Research in Second language Acquisition* (pp. 217-245). Rowley: Newbury.

Lu, C., F.-C. (2001). The acquisition of English articles by Chinese learners. *Second Language Studies, 20*(1), 43-78.

Marcovitch, S., & Lewkowicz, D. J. (2004). U-Shaped Functions: Artifact or Hallmark of Development? *Journal of Cognition and Development*, *5*(1), 113-118.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in Language Acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), i-178.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *TRENDS in Cognitive Sciences*, *6*(11), 465-472.

McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, *11*(2), 113-128.

Murphy, V. A. (2004). Dissociable systems in second language inflectional morphology. *Studies in Second Language Acquisition*, *26*(3), 433-459.

Pinker, S. (1991). Rules of Language. *Science*, *253*, 530-535.

Pinker, S. (1998). Words and Rules. *Lingua*, *106*, 219-242.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73-193.

Pinker, S., & Prince, A. (1991). Regular and Irregular Morphology and the Psychological Status of Rules of Grammar. *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistic Society: General Session and Parasession of The Grammar of Event Structure* (pp. 230-251). Berkeley: Berkeley Linguistic Society.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *TRENDS in Cognitive Sciences*, *6*(11), 456-463.

Pliatsikas, C., & Marinis, T. (2013). Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition*, *16*, 167–182.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43-102.

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*, 156.

Prasada, S., Pinker, S., & Snyder, W. (1990). Some evidence that irregular forms are retrieved from memory but regular forms are rule generated. *Bulletin of the Psychonomic Society*, *28*(6), 519-519.

Rumelhart, D., & McClelland, J. L. (1986). On learning the past tense of English verbs. Implicit rules or parallel distributed processing? In J. L. McClelland, D. Rumelhart, & the PDP Research Group (Eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge: MIT Press.

Say, T., & Clahsen, H. (2002) Words, Rules and Stems in the Italian Mental Lexicon. In Nooteboom S., Weerman F., Wijnen F. (Eds.), *Storage and Computation in the Language Faculty. Studies in Theoretical Psycholinguistics* (Vol 30). Dordrecht: Springer.

Seidenberg, M., & Bruck, M. (1990). *Consistency Effects in the Generation of Past tense Morphology*. Paper presented at the thirty-first meeting of the Psychonomic Society, New Orleans.

Shirai, Y. (1990). *U-shaped Behavior in L2 Acquisition*. In H. Burmeister, & P. L. Rounds (Eds.), *variability in Second Language Acquisition: Proceedings of the Tenth Meeting of the Second Language Research Forum* (Vol. 2, pp. 685-700). Eugene: Department of Linguistics, University of Oregon.

Shirai, Y. (2019). *Connectionism and Second Language Acquisition*. New York: Routledge.

Sivertzen, T. G. (2013). *The role of input in early second language acquisition: A study of the long-term effects of initial extra English input on Norwegian 4th graders' receptive vocabulary*. Unpublished MA thesis. Department of Language and Literature, Norwegian University of Science and Technology.

Slabakova, R. (2008). *Meaning in the Second Language*. Berlin: Mouton de Gruyter.

Slabakova, R. (2013). What is easy and what is hard to acquire in a second language: A generative perspective. In: García Mayo, M, Gutiérrez Mangado, & MJ, Martínez-Adrián, M. (Eds.), *Contemporary approaches to second language acquisition* (pp.5–28). Amsterdam/Philadelphia, PA: John Benjamins.

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

Utdanningsdirektoratet. (2013). *Læreplan i engelsk* (ENG1-03). Retrieved September 17, 2019 from https://www.udir.no/kl06/ENG1-03.

Vulchanova, M., Foyn, C. H., Nilsen, R. A., & Sigmundsson, H. (2014). Links between phonological memory, first language competence and second language competence in 10-year-old children. *Learning and Individual Differences*, *35*, 87-95.

Vulchanova, M., Talcott, J. B., Vulchanov, V., Stankova, M., & Eshuis, H. (2012). Morphology in autism spectrum disorders: Local processing bias and language. *Cognitive Neuropsychology*, *29*(7-8), 584-600.

Vulchanova, M., Baggio, G., Cangelosi, A., & Smith, L. (2017). Editorial: Language Development in the Digital Age. *Frontiers in Human Neuroscience*, *11*, 447. DOI=10.3389/fnhum.2017.00447

Wode, H. (1978). Developmental sequences in naturalistic L2 acquisition. In E. Hatch (Ed.), *Second Language Acquisition* (pp.101-117). Rowley: Newbury.

Appendix

Part 1: Background and participant information

1. Please submit your participant ID for this survey (For instance: ID1498):

Answer:_____

2. When were you born?

Answer:_____

3. Please specify your gender

- o Male
- o Female

4. What is your native language (morsmål)?

- o Norwegian
- o English
- o Other

Please specify your other native language(s):

Answer:_____

5. What other language(s) do you speak in addition to your native language(s)? Please specify language(s) and competence (high – medium – low) in the textbox below.

If you do not speak any additional languages, please write 'none'.

Examples of how to answer: English (medium) Arabic (high) Mandarin/ Chinese (low)

Answer:_____

6. Do you have a diagnosis that could potentially affect your language learning (e.g. dyslexia, impaired hearing, etc.)? If yes, please specify. If no, simply write :'no'

Answer:_____

7. Have you lived in an English speaking country for more than 2 months? If so, where and for how long? If no, simply write: 'no'.

Answer:_____

8. Do you have close family and/or friends that speak English?

Answer:_____

9. Approximately how many hours do you use English throughout a day? (Including at school, and other activities such as reading English online media etc.)

Answer:_____

10. When do you use English in everyday life?

- o While watching TV/movies/series with subtitles
- o While watching TV/movies/series without subtitles
- o When talking to friends/family
- o At school
- o When reading/writing English blogs/websites
- o While reading English books (not including what you read at school)
- o When writing English (not including assignments at school)
- o While playing games (PC-games, X-box, Playstation etc.)

11. Are you involved in one or more activities on your free time that involves use of English? (e.g. theatre/role playing, gaming, etc.)

- o Yes, please specify below
- o No

Please specify which activities:_____

12. How would you rate your own English competence?

- I can understand English text well
- I can understand English speech well
- I can write English well

13. Please rate how well you know English on a scale from 1-5. 1 = Poor, 5 = Very good

	1	2	3	4	5
English:	□	□	□	□	□

Part 2: Tasks

Task 1:

SET: Every year, this guy sets out on a new journey.

What did he do last year?

He_____ out on a journey.

Task 2:

RUN: Joey likes to run 10km every day.

What did Joey do yesterday?

He_____10km.

Task 3:

DO: Joey does something stupid every week.

What did Joey most likely do last week?

He_____something stupid.

Task 4:

MAKE: Betty makes good food when Joey comes over for dinner.

Yesterday, Joey came over for dinner. What did Betty do?

She_____good food for Joey.

Task 5:

GET: Every year, Joey gets Betty something nice for her birthday.

This year, Joey_____Betty a nice present.

Task 6:

SAY: Whenever Joey and Betty argue, they say mean things to each other.

Last time they argued, both_____mean things to each other.

Task 7:

GO: Paul likes to go to the arcade to play games.

Paul_____last Friday to the arcade at 12 o'clock.

Task 8:

TAKE: Paul takes Betty's sister Lisa to go for a hike in the weekends.

Last week, Paul_____Lisa for a hike around a lake for a change.

Task 9:

SEE: Paul loves to see movies on his spare time.

Last night, Paul_____Mad Max: Fury Road at home.

Task 10:

KNOW: Since Paul has seen the movie before, he knows what will happen.

One could say that Paul_____what would happen at the end of the movie.

Task 11:

FIND: Betty's friend Anne likes to find nice stones along the beach.

Once, walking along the beach, Anne_____a green stone!

Task 12:

COME: Anne usually comes over to Joey's birthday parties.

This time, Anne_____late to the party.

Task 13:

THINK: Joey thinks that begin late is rude.

During yesterday's party, when Anna was late, Joey_____that Anne was rude.

Task 14:

GIVE: Anna usually does not give presents.

At the party though, Anna_____Joey a gift!

Task 15:

KEEP: Joey like to keep his presents hidden away from people after parties.

After yesterday's party, Joey_____his presents hidden.

Task 16:

SHOW: Betty like to show pictures of Joey at parties.

At yesterday's party, Betty had_____all her pictures of Joey!

Task 17:

FEEL: Usually, Joey feels happy when Betty shows people pictures of him.

However, at the party, Joey_____embarrassed by some of the pictures.

Task 18:

LEAVE: Most of the time, Lisa leaves early when going to parties.

At yesterday's party, Lisa_____early.

Task 19:

TELL: Paul likes to tell jokes at parties.

At yesterday's party, Paul_____a funny joke that made everybody laugh!

Task 20:

PUT: Betty likes to put berries on top of cakes.

Betty_____berries on top of the cake she baked last year.

Task 21:

HIDE: During hide and seek, you should hide from the others who are playing the game.

Lisa_____an hour ago, and the others have still not been able to find her!

Task 22:

SLEEP: Sleeping is important

However, when you have_____for 15 hours, that's too long!

Task 23:

SING: Betty likes to sing.

Betty_____on TV during last year's American Idol!

Task 24:

BLOW: Joey thinks it is hard to blow out all the candles on the cake.

This year though, Joey_____out all the lights on the cake!

Task 25:

STEAL: Paul likes to steal things.

However, during the party, Paul_____Lisa's heart.

Task 26:

HURT: Paul hates to get hurt when playing sport.

Last time Paul played American football, Paul_____his leg.

Task 27:

WAKE: Joey wakes up 7 o'clock every morning.

However, Joey_____up at 1 o'clock last Sunday!

Task 28:

SHUT: When Paul enters his home, he usually shuts the door when he goes inside.

Yesterday, when Paul came home, he_____the door after himself.

Task 29:

SHAKE: Joey likes to shake his milkshakes before he drinks them.

Yesterday, Joey_____his milkshake more than usual!

Task 30:

SLIDE: It's easy to slide on an icy road and fall over.

Last year, Anna_____on the ice and broke her leg!

Task 31:

SWIM: Betty likes to swim in the weekends.

Last weekend, Betty_____5km in a pool!

Task 32:

TEAR: Every Monday, Paul has to tear down all the posters on the poster-wall.

Last Monday, Paul_____down 100 posters!

Task 33:

FREEZE: Brian likes to prepare all his food a week in advance, and freeze it afterwards.

Last night, after preparing his food, Brian_____his food.

Task 34:

SHINE: Brian likes it when the sun shines on him.

Yesterday, the sun_____through the clouds on Brian.

Task 35:

BITE: Brian is afraid of getting bitten by zombies.

During a nightmare, Brian got_____by a zombie!

Task 36:

BET: Brian has a gambling problem and likes to bet way too much money on poker.

Yesterday, Brian_____10 000kr on a poker game and lost!

Task 37:

FLY: In his spare time, Brian likes to fly.

Yesterday at the party, Brian_____through the window.

Task 38:

LEND: People usually lend Brian money after he loses a poker game.

Joey_____Brian 10 000kr after his last poker game.

Task 39:

SWING: Joey likes to play baseball and swing the baseball bat.

Yesterday, Joey_____the bat and hit a homerun!

Task 40:

CREEP: Kevin likes to creep around the forest.

Last year, Kevin_____around the forest and got in trouble afterwards.

Contents:

Universidade Vigo