Femke Zandee

6270050

Bachelor Thesis

20-7-2020

The Acquisition of Sarcasm in L2 British English After Explicit Instruction

Acknowledgements

This thesis was finished in August 2020, which means that it was written during a turbulent time. It truly would not have been here had it not been for the help of my supervisors, co-supervisor, and family. Therefore, I would like to take some time to thank these people for their help.

First and foremost, I'd like to thank Dr. Aoju Chen, who continually nudged me in the right direction, but always gave me the time to discover the right direction myself. It took many meetings and revisions to get to the finished project, but her assistance truly made this thesis what it is now.

I'd also like to thank dr. Stella Gryllia, for her feedback on my thesis and for aiding in the grading process.

A very special thank you goes out to Nelleke Janssen for helping me with all the technicalities of the results and for providing invaluable aid in figuring out how to use RStudio, as I had never touched the programme before starting this thesis, and now feel comfortable using it, which, considering my skills using other coding programmes, is truly a miracle.

Lastly, I'd like to thank my family, as they supported me throughout the lockdown and kept distractions away from me as much as they possibly could.

Abstract

This study looked at the improvements made by native Dutch L2 speakers of English in their production of sarcastic prosody after a short training by analysing the differences in the prosodic cues before and after the training. Additionally, it aimed to uncover which of these cues is altered first. 12 participants were recorded before and after a short training, after which the data was annotated in Praat and the data separated by cue was extracted using ProsodyPro. The analysis of these results shows that the differences in prosodic cues depended on utterance type and gender. The training helps the participants to sound more like native speakers of English, by moving the production of prosodic cues more to native production values.

Table of Contents

1. Introduction

Sarcasm, a common communicative phenomenon, is a rhetorical device that refers to a process in which speakers use words with the intent to express the direct opposite of the literal meaning of these words. Because the true or intended meaning from these utterances cannot be retrieved from the wording, contextual and prosodic cues are necessary to convey the difference between sarcasm and sincerity. Contextual discrepancy is often enough for a native speaker to recognise a sentence as being sarcastic (Capelli, Nakagawa & Madden, 1990; Creusere, 1999), while prosodic cues aid this recognition (Gibbs & O'Brien, 1991). However, when these contextual cues are not readily available, prosody becomes critical and, for most native speakers, sufficient to make the distinction (Rockwell, 2000; Cheang & Pell, 2013).

However, while universal cues to sarcasm have been theorized, such as a decreased speech rate and increased intensity (Attardo et al., 2003; Rockwell, 2000; 2007), prosodic cues to sarcasm differ among languages. Research has found that the perception of cues is difficult in a second language . Non-native speakers of Cantonese and English did not correctly identify sentences as being sarcastic, while they did identify them in their native language (Cheang and Pell, 2013). Chen and de Jong (2015) added onto this idea by recording native speakers of Dutch while speaking sarcastically in their L2, British English. They concluded that Dutch learners did not appear to have expressed the prosodic cues of sarcasm in British English correctly.

Additionally, like Cheang and Pell (2013), Chen and de Jong (2015) argued that sarcasm could not be produced properly without training. Smorenburg et al. (2015) researched the effectiveness of the explicit training by replicating Chen and de Jong's perception experiments after giving native speakers of Dutch training in British English sarcasm. The participants were asked to, among other things that will be elaborated upon later in this study, record sarcastic utterances in their L2 before and after the focused training. Smorenburg et al.

found that the Dutch learners of English sounded more sarcastic to native speakers of English after the training than before the training (Smorenburg et al., 2015).

Despite that the training has improved the production of sarcasm in L2, it was not clear how differently the Dutch learners of English have used their prosody after the training, compared to their prosody before the training. The present study will attempt to fill this gap by examining the differences between the production of prosodic cues commonly associated with sarcasm before the training and that after the training in Smorenburg et al.'s (2015) study. Additionally, we aim to uncover which cues are altered first, and whether the conditions of utterance type and gender affect the changes made.

In the next section, an overview of the previous studies shall be provided. This will include a background on the definition of sarcasm, how it works in different languages, and a more focussed look at both English and Dutch. In sections 2.2 to 2.6 this background will be elaborated on with the effects of transfer, training, and an in depth look at the effects of gender and utterance type on production. Section 3 will present the research question and hypothesis. The next sections, 4, 5 and 6, contain the methodology of the research followed by the results and the discussion section, which provides suggestions for further research as well. After this, in section 7, the conclusions are given.

2. Theoretical Background

2.1 Sarcasm

The definition of sarcasm varies across sources. The Oxford English Dictionary defines it as a "sharp, bitter or cutting expression or remark", yet in many research papers it is classified as a type of irony (Cheang & Pell, 2008; Rockwell, 2000; Riloff et al, 2013;) or, alternatively, seen as being interchangeable with the term irony (Attardo et al, 2003). On the other hand, Anolli et al. researched vocal patterns and define sarcasm as something that does not necessarily have negative intentions, instead opting to recognise four sub-categories of

sarcasm. These groups are separated between cooperation, which features kind irony and kind banter, and conflict, with sarcastic irony and sarcastic banter. The difference in banter and irony lies in the level of involvement from the speaker, i.e., does the speaker aim to directly criticize or praise the target or to attenuate. Thus, sarcastic banter can be defined as an 'intention to attenuate the critic' (Anolli, Ciceri & Infantino, 2002). The current study, as it is based on Smorenburg et al.'s (2015) training which focussed on sarcastic banter, shall adhere to this focus on sarcastic banter.

Because sarcasm is often claimed to refer to a process in which the intention of the utterance is the literal opposite of the words being used (Kreuz & Glucksberg, 1989; Capelli, Nakagawa & Madden, 1990; Anolli, Ciceri & Infantino, 2002), this intention, as it cannot be inferred from the words used, needs to be made clear through other means. One important aspect is context; for example, saying "It's a beautiful day" during a blizzard. It can be also be conveyed by a body language such as facial expressions (Attardo, Eisterhold, Hay & Poggi, 2003). Additionally, linguistic cues also convey sarcasm. Prosody, for instance plays an important role in expressing sarcasm, as it has been proven that in utterances derived of context, prosodic cues are sufficient for native speakers to recognize the intent (Bryant & Fox Tree, 2005; Rockwell, 2000). Attardo et al. noted that between intonation and body language, intonation is more commonly used to display sarcastic intent (Attardo, Eisterhold, Hay & Poggi, 2003).

There has been quite a lot of research on the prosodic features of sarcasm. Many of these studies cover sarcasm over the complete utterance, contrasting them with sincere sentences and analysing the differences (e.g. Anolli et al., 2002; Cheang & Pell, 2008; Rockwell, 2000). Anolli et al. (2002) compared similar sentences with very extensive context produced by male speakers and found that an ironic tone of voice could be described as using 'caricatured stress'. Cheang and Pell (2008), also using sentences but with no separation of

gender, found that sarcasm could be differentiated from humour, sincerity, and neutrality by a low pitch. Rockwell (2000), who focussed more on the perception, gathered recordings of contextual stories which would feature the same utterance in three intentions; sincerity, sarcasm and posed sarcasm. Participants were able to distinguish the three intentions on vocal cues alone. The vocal cues deemed most important were a slower tempo, a lower pitch level and greater intensity. As such, it seems sarcastic speech does indeed possess different prosodic qualities than sincere speech.

However, not all research has been based on complete utterances. A key word approach was developed that focussed on the target word in a sarcastic utterance (Chen & Boves, 2018), as these key words were deemed to be semantically critical for the sarcastic utterances. For example, in a situation where your lunch has gone stale, your friend might remark "That looks like a tasty lunch." In this instance, the word "tasty" is the target word, as the intended message is that the lunch is the opposite of "tasty". In their study, using key words instead of complete utterances, it was found that key words were produced with a longer duration when pronounced sarcastically, and a lower maximum and minimum pitch, despite the overall pitch span not showing a significant variation. There was also evidence for a gender difference, with women lowering their mean pitch significantly to produce sarcasm, whereas men did not do the same (Chen & Boves, 2018). Similarly, Gonzáles-Fuente et al. (2016) also conducted a study based on singular words, focussing on the final words as well as sentences and found that the prosodic differences were the same at the word and utterance levels.

Most of these studies on the prosody of sarcasm were conducted in English (Rockwell, 2000; Chen & Boves, 2018; Cheang & Pell, 2008). Because of this, we have a generally good idea of how sarcastic speech works in English. It can be assumed that sarcasm in British English is partially marked by changes in the temporal area and cues related to pitch. Time-

related cues for sarcasm are the increased duration of words (Chen & Boves, 2018) and the slower tempo (Rockwell, 2000; Cheang & Pell, 2008). When looking at pitch, the changes made are a lower mean pitch (Cheang & Pell 2008, Rockwell), as well as a smaller pitch range (Cheang & Pell, 2008) and a flatter pitch contour (Chen & Boves, 2018). Chen & Boves, however, did not find significant changes in the pitch range. They did find these changes in the mean pitch, albeit only for women (Chen & Boves, 2018).

However, research on the prosodic qualities of sarcasm has also been conducted in a wide array of other languages. Anolli et al. (2002), for example, found that in Italian pitch was generally raised and the pitch range exaggerated when speaking sarcastically, as opposed to the flattened pitch in English. Yet, similarly to English, the native speakers of Italian also lowered their tempo to convey sarcasm (Anolli, Ciceri & Infantino, 2002). French and Cantonese displayed similar patterns, with a higher mean pitch and pitch range, and a longer duration (Gonzales-Fuente, 2016; Cheang and Pell, 2011). German speakers behave more like English speakers, with a lowered mean pitch and pitch range as well as longer durations to speak sarcastically (Niebuhr, 2014). Mexican Spanish speakers also displayed a smaller pitch range (Rao, 2013). Jansen (2019) found that in Dutch, the effect of pitch is less clear, as the alterations made to express sarcasm varied largely between speakers. However, the pitch range did seem to be increased as well as the duration. Despite the limitation of having few studies per language, the studies showcase similarities as well as differences in these other languages.

Jansen (2019) also compared her results on how sarcasm functions in Dutch to the known prosodic cues for sarcasm in English. While the slower speech rate coincides with the English prosody, it is also mentioned that the insignificant changes in mean pitch and the pitch maximum and minimum are an irregularity when compared to other Germanic languages such as English, where pitch is often lowered. Pitch span is increased when

conveying sarcasm in Dutch but only by female speakers, which is similar to French and Italian but not to English, where pitch span is narrowed. As such, it becomes clear that Dutch differs from English in the use of pitch-related cues to express sarcasm, but the two languages are similar in their use of temporal cues.

2.2 Transfer

As mentioned before, prosodic cues aid the recognition of sarcasm. However, these cues do not seem to work the same in different languages. Cheang and Pell (2013) asked native speakers of both English and Cantonese to listen to sarcastically uttered sentences in their native languages and in a non-native language, i.e., English for the native speakers of Cantonese and vice versa. While the participants had no issues recognising sarcasm in their own languages, sarcasm in their L2 proved more difficult as they frequently marked sarcastic sentences as being sincere (Cheang & Pell, 2013). Thus, language barriers likely obstruct the perception of sarcasm. Production of sarcasm undergoes similar effects. Chen and de Jong proved in their research that L2 learners of English could not efficiently utilise prosody to convey sarcasm. Native Dutch participants recorded sarcastic and sincere utterances in their L2, British English, which were in turn rated by both native speakers of British English and native speakers of Dutch. The speakers from the production task sounded notably less sarcastic to native speakers of English than they did to native speakers of Dutch, suggesting that the Dutch learners of English were unable to express sarcasm effectively in their L2 (Chen & de Jong, 2015).

These issues could have arisen from prosodic transfer from the L1 to the L2. Many studies of prosody in L2 have concluded that L1 transfer is an important factor for the acquisition of both perception and production of prosodic cues in the L2 (Rasier & Hiligsmann, 2007), and that learners of a new language commonly transfer prosodic features from their L1 to their L2, including but not limited to pitch realisation (Li & Post, 2014).

Ueyema (2000) also stated in their research on native speakers of English and Japanese that they transferred prosodic patterns from their L1 to their L2, in these cases, Japanese and English respectively (Ueyema, 2000). Pennington and Ellis (2000) also claimed that less advanced L2 speakers filled up the gaps in their prosodic knowledge with more general prosodic knowledge or contextual cues. Additionally, speakers who had not attained full competence by adulthood chose to transfer their L1 prosody, which they have more experience with, to their L2 (Pennington & Ellis, 2000).

These issues are, however, not limited to beginning learners of an L2. Jun and Oh (2000), for example, researched the prosodic acquisition of Korean by native American English speakers, and found that while advanced speakers had improved the production of pitch boundaries, other surface tones were equally difficult for advanced and beginning speakers of L2 speakers. Additionally, the perception of accentual features of semantic phrases were also difficult for both groups of participants, with *wh*-utterances, utterances that start with the word "what" but are not questions. Advanced speakers only marked these sentence types correctly 63% of the time. Chen and de Jong (2015) separated their participants based on their fluency, creating an intermediate and an advanced group of speakers. The intermediate speakers, who had a B1-B2 level of English, scored significantly lower than the advanced group of speakers. However, on a scale of 1-5, the sarcasm of the advanced group was still only at 2.7 out of 5, while the intermediate group scored 2.43. However, it does suggest that the advanced group's increased exposure aided them in their production (Chen and de Jong, 2015).

This last idea is in line with Cheang and Pell's (2013) theory that experience with the non-native language is necessary for recognition of sarcastic intent, as they claim the issues stem from the differences between languages. In their research, the accuracy of the participants was much greater in their native language than when judging the utterances in a

language more foreign to them. As such, they claimed that '(native) experience' was necessary to recognise sarcasm. Chen and de Jong (2015) hypothesised that the improved production of advanced L2 speakers was due to their experience with the English language, but aptitude might also have an effect. The differences could not result from increased instruction, as prosody of a foreign language is usually not explicitly taught in class. Thus, experience greatly improves the results on both the perception and production of sarcasm.

2.3 Training

Because experience improves these results, the question arises: if experience is necessary for the successful production and perception of the prosody of sarcasm, and prosodic knowledge is not generally taught explicitly, does explicit teaching help in the area of successful utilisation of prosody? Smorenburg, Rodd and Chen (2015) aimed to answer this question and conducted an experiment on the effectiveness of such explicit teaching. The participants, advanced native speakers of Dutch, were asked to produce sarcastic utterances in their L2, British English, before and after a training. The utterances produced before and after the training were rated by native speakers of English on how sarcastic they were perceived to be, and the utterances in the post-test were deemed significantly more sarcastic, with the rating out of 5 increasing from a mean of 2.69 to a mean rating of 3.08. As the participants were all advanced learners of English, which matches the 2.7 rating advanced learners scored in the research of Chen & de Jong (2015), prosodic training can be considered a useful tool for this type of learners, as the training did indeed help.

However, Smorenburg et al.'s (2015) conclusions are only based on the perceptual ratings of native speakers of English. An analysis of the changes in prosody before and after the training has not yet been conducted. It has been discussed previously how a difference in prosodic cues proves to be a hindrance in the production and recognition of the prosody of sarcasm in an L2. Yet, after a relatively short training of approximately two hours, the rating

by the native speakers increased significantly. Because of this, examining which cues were used differently after the training conducted by Smorenburg et al. could help us in discovering which of these prosodic cues are deemed as important for the perception of sarcasm. The current study aims to uncover more details about the changes made in production after an explicit prosodic training.

2. 4 The order of acquisition

There has been limited research on the acquisition of L2 English prosody. As such, it is difficult to construct a hypothesis on the order in which the participants of this study will improve their prosodic production. However, the proven effects of transfer might shed some light on this issue when combined with the effects of proficiency. Li and Post (2014) proved in their research that German learners of English transferred their L1 duration knowledge to English, but the deviations from English could not entirely be explained by it.

Pitch acquisition, however, seems to be slightly different. A study conducted with German learners of English found that in three measures of pitch range, the speakers only produced language-appropriate values in one of the measures. The participants produced values somewhere in between the two languages in the second and third measures, with the results of the second measure being closer to the native English values than the third. While the effect of proficiency was not measured in this research, these relatively positive results were attributed to the advanced proficiency of the participants by the authors (Mennen, Schaeffler & Dickie, 2014). Other research that focussed on English as the L1 with different L2's found that both advanced and beginning speakers had troubles with pitch related prosody such as surface tones (Jun and Oh, 2000), and that those who learn an L2 as an adult have less control over their production of pitch accents (Huang and Jun, 2011).

Based on the aforementioned findings alone, it is impossible to determine whether duration or pitch will be acquired earlier in an L2. It has been shown by other studies that an

L1-independent pattern of the acquisition of pitch and duration does not seem to exist: in Gu and Chen's research, both remained difficult for advanced speakers (2014). To then form a hypothesis on the acquisition of both pitch and duration in L2 English produced by native speakers of Dutch, it is important to acknowledge the transfer-aspect: duration in sarcasm is increased in both English and Dutch and might thus already be more similar in both languages, which could lead to less room for improvement. Pitch related cues, in comparison, do differ between sarcasm in English and Dutch, which would likely cause more negative transfer, but allow for more improvement. As such, the effect of L1 transfer would likely result in duration to be acquired correctly first, due to no positive transfer.

2.5 Utterance Types

Aside from prosodic transfer and proficiency, utterance type also plays a role in prosodic expression of sarcasm. Previous research has shown that utterance type affects how sarcasm is realised prosodically in both English (Chen & Boves, 2018) and Dutch (Jansen, 2019). It also proved to create a difference for the perception of sarcasm produced by L2 speakers of English (Chen & de Jong, 2015). Additionally, in Smorenburg et al.'s (2015) research, the improvement made by participants differed per sentence type, as there are significant differences in the amount of improvement visible in each sentence type. Three different sentence types were used in this research: what-exclamatives, tag questions and declaratives.

As mentioned before, the production of sarcasm in both Dutch and English differed based on utterance type. What-exclamatives, which will be referred to as *wh*-exclamatives for the remainder of this thesis, displayed the largest difference in minimum pitch in English, and were the only sentence type to show significant lowering of minimum pitch in Dutch. Additionally, they also experienced the largest increase of duration in Dutch. Tag-questions underwent similar effects in English; however, in Dutch, it was unlikely that pitch values were altered to make these utterances sarcastic. Declaratives in English featured the lowest

amount of minimum pitch lowering out of the three utterance types in English and were the only type that did not show a significant difference in maximum pitch between sincere and sarcastic conditions. In Dutch, however, maximum pitch was raised, indicating an upwards alteration between sarcasm and sincerity (Chen & Boves, 2018; Jansen, 2019).

The effect on perception was that some sentences were ranked as being more sarcastic than others. While Chen and de Jong (2015) found a significant difference in rating when looking at particle sentences and exclamatives as compared to the other utterance types, no significant difference was found in the rating of declarative sentences and tag questions, the two utterance types from their research this study is interested in (Chen & De Jong, 2015). However, Smorenburg et al. (2015) did find such a difference: both declaratives and *wh*-exclamatives were rated as being significantly more sarcastic than tag-questions in the pre-test. No significant difference between the average rating of declaratives and *wh*-exclamatives could be found (Smorenburg et al. 2015).

Acquisition was also affected by utterance type. The amount of improvement differed significantly between the types: declaratives improved the most, followed by *wh*-exclamatives. Tag-questions improved the least. Additionally, the average results differed significantly between all types after the training. Declaratives were rated higher than *wh*-exclamatives, which were in turn rated higher than tag-questions. After training, the amount of improvement in *wh*-exclamatives ranked second (Smorenburg et al., 2015).

2.6 Gender

Another aspect that affects the production of prosody is gender. Research on sarcasm in both English and Dutch discovered an effect of gender on various prosodic cues related to sarcastic production (Chen & Boves, 2018, Jansen, 2019). Unfortunately, Smorenburg et al.'s research did not provide results separated by gender, and as such, it is unclear whether a gender difference in the learners could affect the rating of sarcasm produced by L2 learners.

In English, two prosodic cues experienced significant differences between genders: main pitch and duration. Main pitch was lowered in sarcastic speech as compared to sincere speech by women, whereas men did not display such an effect and used similar mean pitch in both conditions. Because this was done only by the female participants, Chen and Boves concluded that the reported lower mean pitch in sarcasm is only applicable to women. Both genders used a longer duration in sarcastic speech. Men, however, displayed a significantly larger duration difference than the female participants. Chen and Boves argued that this might compensate for the lack of mean pitch lowering (Chen and Boves, 2018).

Dutch prosodic cues of sarcasm seem to be produced similarly by both male and female participants, but some differences were also uncovered. For instance, the lack of mean pitch change by male speakers as compared to female speakers could only be found in utterance level and not in key word level. On the key word level, men also made significant changes to their mean pitch between sarcastic sentences and sincere sentences. Women did alter mean pitch and minimum pitch in both utterance and key word level. Women were also found to expand their pitch span in sarcastic conditions, but men did not. Duration corresponded closely to the findings in English, with both women and men utilising a longer duration in sarcastic speech but the difference being much larger in men's speech. It was proposed that this phenomenon in British English might be caused by compensation for the lack of pitch lowering, and Jansen proposes this is also the case in Dutch (Jansen, 2019).

3. Current Research

The current research aims to uncover how L2 learners of English alter their use of prosody to express sarcasm after receiving explicit training. It will do so by comparing the sarcastic speech data produced by participants before and after the training from Smorenburg's et al.'s (2015) research. The research question is thus as follows: How do L2 learners of English alter their prosodic production of sarcastic speech after explicit training?

To answer this question, we will look at the changes made in production between the pre-test and the post-test. These changes will additionally be separated by gender and utterance type, to gain a more accurate insight in how these factors affect the acquisition of L2 sarcastic prosody. After the changes have been analysed, they will be linked to native Dutch and native English prosody of sarcasm, to see how these changes have improved the prosody of the participants to sound more like native English.

Our hypothesis is that L2 learners make different changes in their prosody based on the utterance type, and that the participant's gender affects these changes as well. Thus, the improvements made to the production of sarcasm to sound more like native English production of sarcasm will vary between participants and utterance type.

Utterance type should affect the production, as the differences and similarities in prosodic cues vary between English and Dutch. Minimum pitch, for example, is not expected to experience large changes in *wh*-exclamatives, as it is already lowered in both English and Dutch production of sarcasm. Declaratives and tag-questions do not have a lowered minimum pitch in Dutch, and as such have more room for improvement in this aspect. Because Smorenburg et al.'s (2015) research on acquisition discovered that declarative utterances improved most after training, it is expected that declarative utterances will feature the most improvement over all cues. As for the effect of gender, we expect that due to the previous findings on gender-related differences in sarcastic speech, men will improve their duration more than women and women will improve their pitch related cues more.

We also expect a larger increase in duration than pitch related cues overall, as based on previous research it appears that duration is easier to learn for L2 learners than pitch related cues. Despite pitch cues having more room for improvement due to larger differences between English and Dutch sarcastic prosody, the short nature of the training will likely cause a preference for a cue that is easier to acquire.

4. Method

4.1 Participants

The participants consisted of 9 women and 3 men (Mean age = 21.3, SD = 1.7), who studied English Language and Culture at Utrecht University (Smorenburg et al. 2015). All of them were in their second or third year of their bachelor programme and had Dutch as their L1. Eleven of the participants specialised in British English and one of them in American English, and all participants were expected to have reached C1 or C2 level of English proficiency on the Common European Framework of References for Languages (Beeker, Canton, Fasoglio & Trimbos, 2010).

4.2 Production data

The data used in this study are the pre- and post-test data gathered by Smorenburg et al. (2015), which consisted of utterances produced by the participants before and after a training on the prosody of sarcasm. Each participant recorded 48 sentences before the training during a simulated telephone conversation task; 24 of which were uttered in a sarcastic condition and 24 in a sincere condition. They were asked to record them as if they were talking to a friend on the phone, and the participants received written instructions which notified them that they had to use prosody to express sarcastic intent. First, 6 practice utterances were recorded at the start of the experiment while the experimenter was still present. After the experimenter had left, the participants were asked to record the other sentences. Each condition featured an equal number of sentences from three utterance types: *wh*-exclamatives, declaratives and tag-questions. Context was provided to increase naturalness, through recordings made by a native speaker of English representing the friend the participants were speaking to.

After recording the first set of sentences, participants in Smorenburg et al.'s research followed a training that consisted of a presentation on the use of prosody to convey sarcasm in English, explanation of the cues and instruction on how to alter them, individual practice

sessions in which the participants were asked to replicate native speakers' production using visual pitch contours in Praat (Boersma & Weenink, 2020), and feedback by the authors on their production. After the training, a new set of 48 sentences were recorded, with the same distribution of utterance types and conditions as the pre-test.

The data from the pre- and post-test was provided in the form of .wav files. Only the sentences spoken in the sarcastic conditions were used, which were different sentences in the pre- and post-test. The recordings were sorted by keywords, as the original participants received their sentences in randomized order. In Smorenburg's et al. (2015) research, no recordings were excluded. However, not all data proved suitable for analysis in Praat (Boersma & Weenink, 2020). Three utterances from the post-test, two declaratives and one *wh*-exclamative, were excluded due to inaudibility or problems encountered in Praat (Boersma & Weenink, 2020).

4.3 Data annotation

The sentences were annotated in Praat (Boersma & Weenink, 2020), with markers for the beginning and end of each sentence and keywords, on separate tiers. In the current study, the focus lies on the keywords, due to the issues that stem from complete sentences often not being lexically identical, as raised by Chen & Boves (2018). The keywords were selected by hand based on which word formed the sarcastic contrast and was thus deemed the target word. For example, in the sentence "She's a healthy lady", the word "healthy" was the designated keyword, as the provided context proved that the lady in question did not lead a healthy lifestyle. Because there were 24 sentences, there were also 24 keywords in the pre- and post-test; 2 of these keywords occurred in both tests.

Using ProsodyPro (Xu, 2013), mean pitch in Hertz, minimum pitch in Hertz, maximum pitch in Hertz and duration in milliseconds were extracted from the keyword-tier.

Pitch span was calculated in Hertz by subtracting the minimum pitch from the maximum pitch.

The effects of the experimental variables on pitch span, minimum pitch, mean pitch, maximum pitch and duration were analysed using linear mixed-effect modelling in R. To do so, the lme4 package was used (Bates, Maechler, Bolker & Walker, 2015). The fixed factors were TEST PHASE: (pre-test vs. post-test), UTTERANCE-TYPE (*wh*-exclamative, declarative or tag-question) and GENDER (male vs. female). The random factors were PARTICIPANT and ITEM, i.e., the keywords. A model was built up for every prosodic cue, adding the fixed factors and their interactions step by step in the order mentioned above and comparing each model to the previous iteration using anova's. Any factor or interaction that did not improve the model was removed, unless it was involved in a new interaction in a next model. Additionally, the average results for all cues during the pre- and post-test were calculated to form the graphs, separated by the fixed factor of the best fitting model.

5. Results

The best-fit linear mixed effect model for each variable is reported. As the effect of TEST PHASE (2 levels: pre-test, post-test), is this study's primary interest, only the significant interactions between TEST PHASE and the other two fixed factors, UTTERANCE TYPE (3 levels: declaratives, exclamative, tags) and GENDER (2 levels: male, female) will be reported. If no such interactions exist, the significant main effect of TEST PHASE will be reported. The interactions between TEST PHASE and the other fixed factors were further analysed with mixed effects models to determine the effect of the TEST PHASE on the specific utterance type and/or gender.

5.1 Mean Pitch

Results from the mixed models showed that there were significant main effects of both TEST PHASE ($p<0.001$) and UTTERANCE TYPE ($p<0.001$) on the mean pitch of the keywords, as well

as a significant interaction between the two ($p<0.001$). This was deemed the best-fit model. Gender did have a significant main effect ($p<0.001$), but no interaction with the test phase could be found ($p=0.119$).

Because the interaction between TEST PHASE and UTTERANCE TYPE was found to be significant, further testing was conducted. Testing on the specific utterance types revealed no significant effect of TEST PHASE on the mean pitch of the keywords in declarative sentences ($p=0.726$). TEST PHASE did have a significant effect on *wh*-exclamatives ($\beta=36.331$, $SE=6.638$, $t(190)=5.474$, $p<0.001$): *wh*-exclamatives were generally produced with a lower pitch in the post-test than in the pre-test, with averages of 206.51 Hz and 220.16 Hz respectively. Additionally, TEST PHASE also had a significant effect on tag-questions ($\beta=13.653$, $SE=6.616$, $t(192)=2.064$, $p=0.041$). In the pre-condition, the average mean pitch was 251.32 Hz, whereas in the post-test it was once again much lower with a mean of 214.36 Hz. Figure 1 illustrates the fall in average mean pitch.

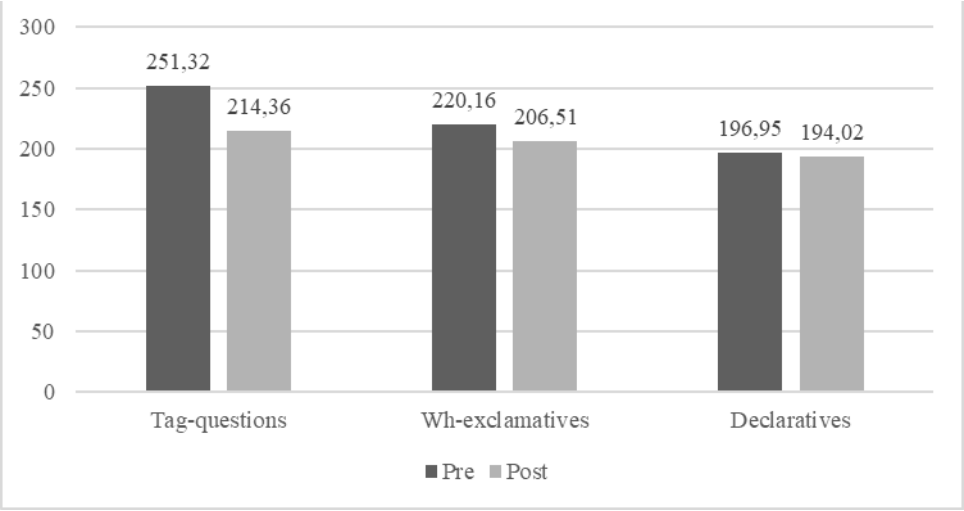

Figure 1. The effect of TEST PHASE on the mean pitch divided by UTTERANCE TYPE

5.2 Minimum Pitch, Maximum Pitch and Pitch Difference

A significant effect of TEST PHASE ($p<0.001$) on the minimum pitch was found, as well as a significant main effect of UTTERANCE TYPE ($p<0.001$), and significant effect of GENDER

(p=0.006). However, no significant interaction was found between condition and UTTERANCE TYPE(p=0.802), nor between TEST PHASE and GENDER (p=0.179). The minimum pitch was lowered between conditions: in the pre-test, the average minimum pitch was 165.38 Hz and the average minimum pitch in the post-test was 154.45 Hz.

Because of the lack of interactions further research on the specific conditions was deemed unnecessary. Additionally, no significant main effect of TEST PHASE was found in either the maximal pitch (p=0.056) nor the pitch span (p=0.974).

5.3 Duration

Results from the mixed models showed significant main effects of TEST PHASE (p<0.001) and UTTERANCE TYPE (p<0.001) and GENDER (p<0.001). A significant interaction between UTTERANCE TYPE and TEST PHASE (p<0.001) was found, as well as an interaction between GENDER and TEST PHASE (p=0.010). The best-fit model thus displayed significant interaction between TEST PHASE and GENDER as well as TEST PHASE and UTTERANCE TYPE.

Further testing displayed a significant main effect of TEST PHASE on all individual utterance types. The keywords in declarative sentences (β=-201.02, SE=29.15, t(190)=-6.897, p<0.001) showed an increase in duration, with the mean of pre-test being 520.62 and that of the post-test being 718.69 milliseconds. Tag-question keywords also underwent a significant increase in duration (β=-97.29, SE=18.20, t(192)=-5.345, p<0.001). The keywords were an average of 412.26 milliseconds in the pre-test, which increased to 509.56 milliseconds in the post-test. The remaining utterance type, the *wh*-exclamative, also showed a significant effect of TEST PHASE (β=117.41, SE=22.56, t(190)=5.205, p<0.001). However, the average duration of the keywords went down instead of up, with 648.18 milliseconds on average in the pre-test, and 530.95 milliseconds in the post-test. The change in average duration in milliseconds is displayed in a Figure 2.

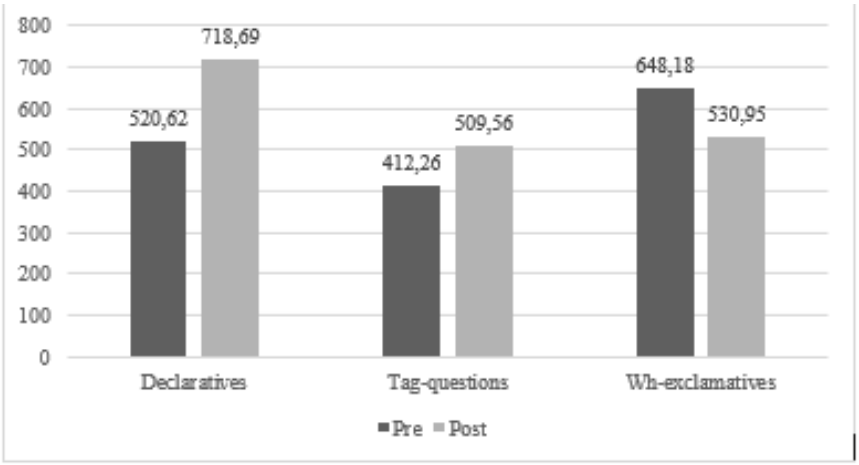

Figure 2. Average Duration in Milliseconds Separated by Utterance Type

Both men and women underwent a significant change in average duration between the two conditions. Men displayed a significant effect of TEST PHASE (β=-122.53, SE=29.10, t(142)=-4, p<0.001) and increased their average duration of the keywords from 496.56 milliseconds in the pre-test to 619.89 milliseconds in the post-test. Women, who also showed a significant effect of TEST PHASE (β=-40.18, SE=17.05, t(430)=-2.357, p=0.019) lengthened their duration as well: in the pre-test, their average duration was 536.40 milliseconds whereas in the post-test it had risen to 574.31 milliseconds. The change in duration separated by gender is displayed in Figure 3.

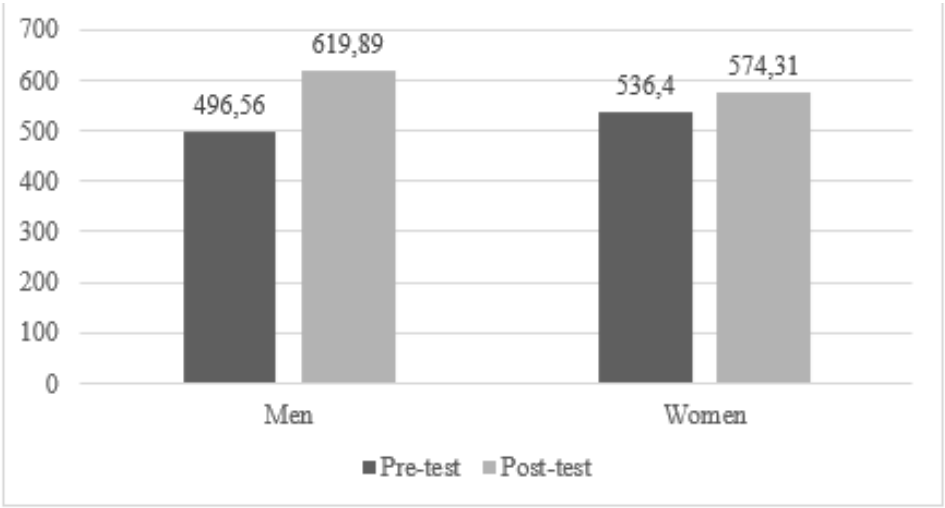

Figure 3. Average Duration in Milliseconds Separated by Gender

6. Discussion

6.1 Pitch-related measurements

We have found that the mean pitch dropped significantly in the post-test, compared to that in the pre-test in two out of three utterance types. Considering that L1 English speakers generally lower their mean pitch when speaking sarcastically, this result suggests the participants succeeded in lowering their mean pitch to sound more sarcastic in *wh*-exclamatives and tag-questions.

A possible explanation for the lack of a decrease in mean pitch in declaratives could be that the average mean test for declaratives in the pre-test was already quite low: the average mean pitch was 196.95 Hz, whereas the averages of mean pitch for the other two utterance types were both above 220 Hz. As such, it is possible that it felt unnatural for the participants to drop their pitch even lower. It could also be caused by some inhibitions stemming from the native language of the participants: in Dutch, mean pitch did not differ in sarcastic and sincere declarative utterances. This could have transferred to sarcasm production in English.

Furthermore, the minimum pitch dropped significantly from the pre-test to the post-test, whereas the maximum pitch and the pitch difference hardly differed between the test-phases. The drop in minimum pitch is compatible with the known changes made in minimum pitch between sincere and sarcastic speech in English.

6.2 Duration

We have found that the duration of the key words changed significantly in the post, compared to pre-test, and the change was larger in male speakers than in female speakers. This indicates that the participants did alter their duration to sound more like native speakers of English while producing sarcastic utterances. While all utterance types displayed great differences in duration between pre- and post-test, as visible in Figure 2, not all of the utterance types showed an increase in duration. Tag-questions and declaratives showed a large lengthening of keywords in the post-test, whereas *wh*-exclamatives displayed a drop in duration larger than

110 milliseconds. This is unexpected, as in both Dutch and English *wh*-exclamatives are lengthened in sarcastic utterances. Since the participants were not instructed to shorten their *wh*-exclamatives in sarcastic utterances during the training, the decrease in average duration is likely caused by other effects.

One possible explanation for these results lies in the difference in keywords between the pre-test and post-test. If the keywords for *wh*-exclamatives were shorter in the post-test on average and the other utterance types did not undergo the same effect, it would explain the results. After evaluating the average number of syllables per utterance type in the pre- ant post-test it became clear that the *wh*-exclamatives did indeed see a decrease in length, with the pre-test keywords having 2,4 syllables on average and the post-test keywords 2. Tag-question keywords averaged on 1,6 syllables in the pre-test and 1,5 in the post-test, whereas the declarative keywords rose from an average of 1,9 syllable per keyword in the pre-test to 2,8 syllables in the post-test. This largely explains and likely caused the results for duration divided by utterance type. Fortunately, the average number of syllables per keyword between the pre-test and the post-test when not divided by utterance type, 2,0 and 2,1 respectively, does not differ enough to cause worry for the effect of test-phase on duration overall. Because the difference between the results from the *wh*-exclamatives and the other two utterance types is rather large, it would be of great importance to eliminate the possible difference in word length between the pre-test and the post-test for any further research.

Furthermore, the larger increase in duration in the male participants is comparable with the production of sarcasm by native speakers of English (Chen & Boves 2018). Following Chen and Boves (2018) and Jansen (20190, we suggest that male speakers seem to utilise duration in compensation for limitation in lowering the mean pitch to sound sarcastic.

6.3 General discussion

The current research's hypothesis was that learners make different changes in their prosody based on the utterance type as well as gender. Because a significant interaction between TEST PHASE and UTTERANCE TYPE was found in both mean pitch and duration, and duration also had a significant interaction between TEST PHASE and GENDER, this hypothesis is borne out. These interactions proved that gender and utterance type affect the acquisition of the prosodic cues of sarcasm.

We expected that participants will alter cues differently in different utterance types, and more specifically, that declarative sentences would improve the most out of all utterance types, because its rating had improved the most. This second part of the hypothesis was proven to be incorrect, as declarative sentences were the only utterance type in which the mean pitch was not significantly changed. The only significant change to declarative sentences that other utterance types did not experience was a larger increase in duration. This implies that this increased duration might be the reason for the improved rating. However, this research does not cover all prosodic cues associated with sarcastic speech, and, as such, other cues that have not been covered might have been the cause for this increase in rating.

Despite this unexpected lack of change in the mean pitch of declarative sentences, cues were still altered differently across utterance types. Mean pitch was lowered in *wh*-exclamatives and tag-questions but showed no significant change in declaratives. Duration was increased in declaratives and tag-questions but decreased in *wh*-exclamatives. While the results for duration were unexpected based on previous research, the differences in mean pitch alteration can be based on the differences between Dutch and English, and thus supports our hypothesis.

Our theory for gender-related differences was that women would improve more on pitch related cues and men more on duration. This theory is partially borne out. Men did indeed improve their duration more than the female participants, and men did not show an

interaction between TEST PHASE and GENDER in pitch related cues. As such, it can be explained by the idea that men put more emphasis on duration to make up for a lack of pitch related changes between sincere and sarcastic speech, which supports our hypothesis. However, the interaction between TEST PHASE and GENDER was not significant either for women, which is contrary to the expectations. While there was an effect of GENDER on mean pitch, this likely resulted from men and women using a different mean pitch in general instead of resulting from sarcastic speech. Thus, the female participants did not perform as expected.

Overall, a significant difference between conditions was found in duration, as well as a significant difference in the mean pitch and minimum pitch. No significant difference was found in maximum pitch, nor in pitch difference. Our theory that only duration would improve due to the positive transfer and the short length of the experiment is thus disproven. However, the fact remains that the change in minimum pitch was relatively small between the pre-test and the post-test, and that no significant changes were made to maximum pitch nor pitch difference, despite maximum pitch being significantly lowered in native English production of sarcasm (Chen & Boves, 2018). Additionally, minimum pitch did not differ by utterance type when produced by the L2 learners of Dutch, but this difference is quite pronounced in native English (Chen & Boves, 2018). This could result from the native Dutch usage of pitch related cues, as minimum pitch and maximum pitch are rarely altered in Dutch, or even raised (Jansen, 2019) As such, it seems likely that pitch cues also suffered from transfer, and that this transfer caused them to be more difficult for the participants to alter.

Finally, we aimed to link these findings to the native Dutch and native English production of sarcastic prosody to gain an insight as to how these changes improve the participants' prosody. As expected, the keywords were indeed produced with a longer duration, matching native English production. Additionally, mean pitch and minimum pitch were both lowered, which is also in line with native English sarcastic speech. While most

utterance-type specific alterations also moved closer to native English, some alterations did the opposite. The decrease of duration in *wh*-exclamatives, for instance, seemingly moves it further away from native English. Male participants performed as hypothesised and thus, closer to how male native speakers of English produce sarcasm, whereas women did not.

6.4 Limitations

There are several limitations in the current study. First, Smorenburg et al.'s research was not based on keywords, but was rated by how the reception of complete utterances by native speakers changed before and after the training. A keyword-based approach was chosen for this study, based on previous research. (Chen & Boves, 2018; Jansen, 2019). The keywords in the sentences used in Smorenburg's et al.(2015) research thus needed to be manually marked and chosen. Because the keywords were not decided beforehand, the sentences in the pre- and post-test were not the same, which resulted in different keywords being compared with one another between conditions. Because of this, further research should feature a new list of keywords, which should be reused in the post-test. Perhaps the words could be used in different sentences within the same utterance types in the post-test, to avoid effects of repetition.

Secondly, the results on the effects of gender are based on a pool of participants which was fairly homogeneous. Only 3 male students participated in Smorenburg et al.'s initial study (2015), compared to 9 female students. The lack of male students could affect the averages in their gender, as one outlier could greatly increase or decrease the number. Future research should, if a focus on gender is deemed interesting, make sure the number of participants of each gender is equal or nearly equal.

Lastly, it could be interesting to conduct a statistical analysis on the production of sarcasm by L2 learners of English to the production by native speakers. The comparisons in this research are based on earlier conclusions made about how native speakers of English alter

their speech to sound sarcastic instead of sincere. Asking native speakers of English to produce the exact same sentences as the L2 speakers of English and comparing the averages of the L2 speakers before and after the training to the averages of the native speakers could prove insightful to how much and what the L2 speakers need to change in order to sound more sarcastic.

7. Conclusion

At the start of this study, we aimed to answer our research question "How do L2 learners of English alter their prosodic production of sarcastic speech after explicit training?". It was hypothesised that the participants would make changes to their prosody depending on utterance type and gender. This hypothesis was proven to be true. The results showed interactions between these two factors and the test phase. Specifically, prosodic cues were altered differently per utterance type, but declarative sentences only showed a significantly larger improvement in duration, and none of the other cues tested. As for gender, it was expected that men would show more improvement in duration and women more improvement in pitch related cues. The male participants did indeed fall into this pattern, showing a larger increase in duration, but women made no discernible larger improvement in pitch related cues.

The study also aimed to connect these findings to how the production of sarcasm functions in native English. Both the increased duration and the lowering of mean pitch and minimum pitch brought the L2 speakers of English closer to native English production of sarcasm, explaining the perceived difference in the degree of sarcasm expressed by Dutch learners of English after brief explicit training in Smorenburg et al. (2015). Additionally, the effects of gender fit into the gender difference in how native speakers of English produce sarcasm. However, while most effects of utterance types moved the speech closer to native-

like, some utterance-specific alterations seemed to move them further away, such as *wh*-exclamatives having a shorter duration in the post-test than in the pre-test.

In conclusion, the prosodic production of sarcasm in L2 English is altered by training mostly by an increase in duration and a decrease in mean and minimum pitch. These alterations move the L2 English speakers' production of sarcasm closer to native English production of sarcasm. The lack of change in maximum pitch and pitch span still separates L2 production from L1 production. Further research should focus on eliminating the limitations mentioned previously in section 6.4, and closely comparing L2 production to L1 production in a more numerical manner could provide valuable insights, for instance, in L2 acquisition of prosody.

References

Anolli, L., Ciceri, R., & Infantino, M. G. (2002). From "blame by praise" to "praise by blame": Analysis of vocal patterns in ironic communication. *International Journal of Psychology*, *37*(5), 266-276.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, *16*(2), 243-260.

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

Beeker, A., Canton, J., Fasoglio, D., Trimbos, B. (2010), Eindtermen havo, vwo. Europees Referentiekader Talen. Nationaal Expertisecentrum leerplanontwikkeling (SLO). Web. http://www.erk.nl/docent/streefniveaus/havo/

Boersma, Paul & Weenink, David (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.09, retrieved 26 January 2020 from http://www.praat.org/

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and speech*, *48*(3), 257-277.

Capelli, C. A., Nakagawa, N., & Madden, C. M. (1990). How children understand sarcasm: The role of context and intonation. *Child Development*, *61*(6), 1824-1841.

Cheang H.S./M.D. Pell, (2013), *Recognizing sarcasm without language: A crosslinguistic study of English and Cantonese*, in S. Attardo/M.M. Wagner/E. Urios-Aparisi (eds.), *Prosody and Humor*, John Benjamins B.V., Amsterdam-Philadelphia, 15-36.

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech communication*, *50*(5), 366-381.

Cheang, H. S., & Pell, M. D. (2011). Recognizing sarcasm without language: A cross-linguistic study of English and Cantonese. *Pragmatics & Cognition*, *19*(2), 203-223.

Chen, A., & Boves, L. (2018). What's in a word: Sounding sarcastic in British English. *Journal of the International Phonetic Association*, *48*(1), 57-76.

Chen, A., de Jong, D., & Chini, M. (2015). Prosodic expression of sarcasm in L2 English. *Marina Chini (ed.) L*, *2*, 27-37.

González Fuente, S., Prieto Vives, P., & Noveck, I. A. (2016). A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French. *Barnes J, Brugos A, Shattuck-Hufnagel S, Veilleux N, editors. Speech Prosody 2016; 2016 May 31-June 3; Boston, United States of America. [place unknown]: International Speech Communication Association; 2016. p. 902-6. DOI: 10.21437/SpeechProsody. 2016-185*.

Gu, Y., & Chen, A. (2014). Information status and L2 prosody: A study of reference maintenance in Chinese learners of Dutch. In J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. Schiller, & E. van Zanten (Eds.), *Above and Beyond the Segments: Experimental linguistics and phonetics*. (pp. 120-130). Amsterdam: John Benjamins Publishing Company. 10.1075/z.189.10gu

Huang, B. H., & Jun, S.-A. (2011). The Effect of Age on the Acquisition of Second Language Prosody. *Language and Speech*, *54*(3), 387-414. https://doi.org/10.1177/0023830911402599

Jansen, N. (2019). Prosodic markers of Dutch sarcasm: Variation related to sentence type, speaker gender, and personality (Master's thesis). University of Utrecht, Utrecht.

Jun, S. A., & Oh, M. (2000). Acquisition of second language intonation. In *Sixth International Conference on Spoken Language Processing*.

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of experimental psychology: General*, *118*(4), 374.

Li, A., & Post, B. (2014). L2 acquisition of prosodic properties of speech rhythm: Evidence from L1 Mandarin and German learners of English. *Studies in Second Language Acquisition*, *36*(2), 223-255.

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.

Niebuhr, O. (2014). "A little more ironic"–Voice quality and segmental reduction differences between sarcastic and neutral utterances. In *Proceedings of the 7th International Conference on Speech Prosody* (pp. 608-612).

Pennington, M. C., & Ellis, N. C. (2000). Cantonese speakers' memory for English sentences with prosodic cues. *The Modern Language Journal*, *84*(3), 372-389.

Rao, R. (2013). Prosodic consequences of sarcasm versus sincerity in Mexican Spanish. *Concentric: Studies in Linguistics*, *39*(2), 33-59.

Rasier, L., & Hiligsmann, P. (2007). Prosodic transfer from L1 to L2. Theoretical and methodological issues. *Nouveaux cahiers de linguistique française*, *28*(2007), 41-66.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 704-714).

Rockwell P., (2007), Vocal features of conversational sarcasm: A comparison of methods. *Journal of Psycholinguistic Research* 36, 361-369. doi: 0.1007/s10936-006-9049-0.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic research*, *29*(5), 483-495.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Smorenburg, Laura, Joe Rodd & Aoju Chen. 2015. The effect of explicit training on the prosodic production of L2 sarcasm by Dutch learners of English. Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015), Glasgow, 1–47.

Ueyama, M. (2000). *Prosodic transfer: an acoustic study of L 2 English vs. L 2 Japanese* (Doctoral dissertation, UCLA).

Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France. 7-10.