Estudios de Fonética Experimental
Journal of Experimental Phonetics
ISSN: 1575-5533 · ISSN-e: 2385-3573

AI is great, isn't it? Tone direction and illocutionary force delivery of tag questions in Amazon's AI NTTS Polly

Alfonso Carlos Rodríguez Fernández-Peña [1] 0000-0002-0273-8442

[1] Universidad de Oviedo (Spain)

DOI: 10.1344/efe-2023-32-227-242

Corresponding address: rodriguezalfonso@uniovi.es

Received: 29/06/2023 Accepted: 03/11/2023 Published: 28/11/2023

Rodríguez Fernández-Peña, A. C. (2023). AI is great, isn't it? Tone direction and illocutionary force delivery of tag questions in Amazon's AI NTTS Polly. *Estudios de Fonética Experimental*, *32*, 227–242. https://doi.org/10.1344/efe-2023-32-227-242

Abstract

This work provides a descriptive analysis of the tone direction and its inherent illocutionary force in question tags delivered by Amazon's neural text-to-speech system *Polly*. We included three types of tag questions (reverse-polarity tags — both positive and negative —, copy tags and command tags) for which 10 sentences were used as input in each case. The data included 600 utterances produced by British and American English voices currently available on Amazon's NTTS. The audio files were examined with the speech analysis software *Praat* to identify the tone pattern for each utterance and confirm the intended illocutionary force. The results show that Amazon's AI speech synthesis technology is not yet fully reliable and produces a high rate of utterances whose pragmatic load is undesired when using natural spontaneous speech traits as question tags.

Keywords

illocutionary force; tag questions; intonation; text-to-speech; artificial intelligence (AI)

La inteligencia artificial es genial, ¿verdad? Dirección del tono y vehiculación de la fuerza ilocucionaria en las preguntas ratificadas de Polly, el sistema de texto a voz AI de Amazon

RESUMEN

Este trabajo ofrece un análisis descriptivo de la dirección tonal y la fuerza ilocucionaria en las preguntas ratificadas (o *tag questions*, en inglés) generadas por el sistema de texto a voz Polly de Amazon. Se examinan tres tipos de preguntas (polaridad revertida — positivas y negativas —, polaridad idéntica y orden) utilizando 10 frases como muestra para cada una. Se emplearon las voces sintéticas disponibles en inglés británico y americano y se generaron un total de 600 muestras de audio. Estas se analizaron con *Praat* para identificar el patrón tonal y confirmar la fuerza ilocucionaria presente en ellas. Los resultados revelan que la tecnología de síntesis de habla de Amazon aún no es completamente fiable, ya que produce un alto número de frases con una carga pragmática inadecuada para lograr una entonación natural en las preguntas ratificadas en inglés.

PALABRAS CLAVE

fuerza ilocucionaria; preguntas ratificadas; entonación; texto a voz; inteligencia artificial (I.A.)

La intel·ligència artificial és genial, oi? Direcció del to i vehiculació de la força il·locutiva en les preguntes amb cua de Polly, el sistema de text a veu AI d'Amazon

RESUM

Aquest treball proporciona una anàlisi descriptiva de la direcció tonal i la força il·locutiva inherent en les preguntes amb cua (o *tag qüestions*, en anglès) generades pel sistema de text a veu Polly d'Amazon. S'han examinat tres tipus de preguntes (de polaritat invertida —tant positives com negatives—, de còpia i d'ordre), utilitzant 10 frases com a mostra de cada cas. Es van utilitzar les veus sintètiques disponibles per a l'anglès britànic i l'americà, i es van generar un total de 600 enunciats. Aquests fitxers d'àudio es van analitzar amb *Praat* per identificar-ne el patró tonal i confirmar la força il·locutiva que s'esperava. Els resultats indiquen que la tecnologia de síntesi de parla d'Amazon encara no és del tot fiable, ja que en el moment de produir trets de la parla espontània natural com les marques de pregunta es produeixen força enunciats amb una càrrega pragmàtica no desitjada.

MOTS CLAU

força il·locutiva; preguntes amb cua; entonació; text a parla; intel·ligència artificial (IA)

1. Introduction

The production of audiovisual material is both widespread and rapidly expanding nowadays. By means of artificial intelligence (AI), some corporations can bring long gone celebrities back to life for profit, and we are witnessing how current Hollywood movie stars are signing away their vocal and physical rights to be used in deepfakes in the future. There is already a plethora of internet multinationals that provide text-to-speech (TTS) services, making synthetic voices a practical possibility. This revolutionary technology can be used for learning purposes, such as e-learning online modules, audio guides, audio books, video games, corporate and commercial voice-overs, IVRs[1], etc. The quality of the voices provided by these companies has also improved with the years, and what started as robotic monotone voices has now reached a level of realism that is sometimes difficult to discern from a real human voice. A great improvement has been achieved at the segmental level with the development of TTS systems, as noted by Cohen et al. (2004, p. 24), Kim et al. (2022, p. 1), Shen et al. (2018, p. 1), and van den Oord et al. (2017, p. 8). However, even though these TTS systems could faithfully reproduce human speech at the segmental level, the delivery of suprasegmentals or prosodic traits, namely stress and intonation, which carry a significant load of illocutionary force and its corresponding pragmatic meaning (Wells, 2006; Mateo, 2014; Gómez González & Sánchez Roura, 2016) is questionable. The aim of this work is to provide a descriptive analysis, both qualitative and quantitative, of the delivery of tag questions and their illocutionary force in English by Amazon's world-famous neural text-to-speech (NTTS) service, *Polly*.

Tag questions are an excellent empirical domain for a study on synthetic speech because they include a diverse array of grammatical elements that contribute to their interpretation. They are syntactically mixed, consisting of a declarative phrase, or anchor, and a shortened interrogative clause, or tag, in a paratactic connection. We argue that this complex language shape is paralleled by a complex discourse function. Moreover, prosody, including intonation, intonational phrasing, and stress, plays a crucial role in computing the discourse function of tag questions. There are several studies that postulate there is a nexus between the final intonational contour of the tag and the discourse function of English intonation (Sadock, 1974; Wells 2006; Parrot, 2010; Mott, 2011; Mateo 2014; Gómez González & Sánchez Roura, 2016; Rodríguez Fernández-Peña, 2022, etc.) Therefore, tag questions offer an attractive testing ground for investigating the pragmatic and illocutionary contributions of intonation and the way a TTS system as Amazon's Polly, which is used by a large number of companies worldwide, produces them.

2. Tag questions and intonation

English intonation is commonly acknowledged to possess a discourse function, often referred to as cohesive (Wells, 2006) and textual structure (Tench, 2009). This function is responsible for providing coherence, comprehensibility, and structure to spoken discourse, similar to how punctuation operates in written language. In a conversation, the discourse function also plays a role in signaling turn taking or indicating when someone has finished speaking. It enables listeners to determine whether information is relevant, overlapping with what is referred to as *accentual function*.

Both tonicity (or the position of the nuclear accent) and tones are used in English to indicate which information is relevant in an utterance. The rule of deaccenting given information, or anaphora rule (Mott, 2011, p. 205) signals which part of the utterance shall be taken as new/relevant and which as old. This is illustrated in example (1).

(1) No woman had ever made that step from royal mistress to the throne, | getting the \Queen, | a \real Queen, | out of the way.

[1] IVR stands for Interactive Voice Response and is used as a computer-operated telephone system.

As regards to tones, rising tones indicate non-finality and imply that information is sought, or anticipated, rather than unloaded, whereas falling tones suggest the opposite, that is, finality and allow the hearer to understand that no more new information is coming (Collins & Mees, 2013, p.147).

(2) We don't know to what extent she /loved him, | if she ever /did, | or if she operated on a basis of cold am\bition.

In example 2, the rising tones allow the hearer realise that additional information is forthcoming, until the falling tone is perceived indicating the end of the message.

Another case in which intonation plays a crucial role in the interpretation of a message is observed in tag questions. Tag questions are frequently used in spontaneous oral speech (Leech & Svartvik, 1994) and are syntactic structures consisting of only an auxiliary verb and a pronoun, added at the end of a statement in speech and sometimes in informal writing (Swan, 2005, p. 469). What is interesting about these constructions, as pointed out by Gómez González and Sánchez Roura (2016, p. 308), is that their intonation is "pragmatically determined" since they may encode two different illocutionary forces, functioning either as a genuine question or as a request for confirmation on the information provided in the statement. For these scholars, "intonation, then plays a disambiguating role, conveying the pragmatic force of the tag" (ibid.).

Tag questions are often divided into two branches, each having two subcategories, associated with two distinct intonation patterns, seemingly conveying their own illocutionary force (Kay, 2006; Gómez González & Sánchez Roura, 2016).

The first group includes balanced tags (Collins & Mees, 2013, p. 152), also known as reverse polarity tags (Cruttenden, 2014, p. 95), that is, the tag is negative if the main clause is positive, and vice versa. Balanced tags have their own IP and can have different communicative meanings depending on the tone they have (rising or falling). If the tag is not a real question, and one is sure of the answer, then the tone is a fall [\], and the receiver should not answer the question. Falling tags are tricky because they can have different meanings. They are not real questions and, instead of seeking information, they might aim to elicit agreement (Thomson & Martinet, 1986, p. 80; Swan, 2005, p. 88; Wells, 2006, p. 49). Thus, an answer or confirmation may be provided. However, they can also be used to express an opinion (Leech & Svartvik, 1994, p. 151), in which case, no answer is expected, since "the sentence is more like a statement than a question" (Leech & Svartvik, 1994, p. 127).

On the other hand, tags with a rising intonation [/] — a yes-no rise — sound less assertive (Kay, 2006, p. 694) and normally imply a question. They are genuine questions (Vince & Emmerson, 2003, p. 182), so the receiver should answer the tag. Example 3 includes instances of reverse polarity tags, both positive and negative.

(3) a. The match was a disaster, | /wasn't it?
(rising = asking, answer expected)
b. The match was a disaster, | \wasn't it?
(falling = confirmation, answer not expected)

Moreover, within reverse polarity tags, we have what McCawley (1988) labelled as fake negative tags. These constructions "superficially have negation in the host (and not in the tag), but the host is nonetheless a positive polarity environment" (Kay, 2006, p. 694). Their main characteristics are a rising tone for the tag and the illocutionary force of a timid suggestion, as shown in example 4 (Kay, 2006, p. 694).

(4) Example 1
a. You wouldn't rather go to the \movies, | /would you?
b. *I wouldn't rather go to the movies.

The other group consists of copy tags (Gómez González & Sánchez Roura, 2016, p. 308), also known as unbalanced tags (Collins & Mees, 2013, p. 152), comment tags (Thomson & Martinet, 1986, p. 80) or constant-polarity tags (Wells, 2006, p. 49;

Cruttenden, 2014, p. 296), which are commonly used to show surprise or disbelief, even sarcasm (Cattell, 1973, p. 612, citing Lakoff, 1969). They have the same polarity as the main clause (normally positive-positive) and are always accompanied by a rising tone (Kay, 2009, p. 694; Wells, 2006, p. 49; Parrot, 2010, p. 116), or a low rise[2] according to Cruttenden (2014, p. 296). Tench (2009) claims that they do not necessarily need their own separate intonation phrase (IP). Moreover, Cruttenden also suggests that "falling tones are impossible" (2014, p. 296) for this type of tag. In terms of illocutionary force, Cattell (1973, p. 615) considers that the same polarity tags usually accompany sentences that the speaker is not putting forward as his own but is "citing in order to ask the listener if it is his". According to Kay (2006, p. 694) these kinds of tags can appear in utterances conveying either belligerence or docility, as illustrated in example 5:

(5) JOHN: So you're happy with the promotion, | /are you? *(He thought Sam would not like it because the workload is much heavier.)*
SAM: Sure, the new post is truly challenging. I love it!

Tags can also be attached to other types of clauses, apart from statements, which are more restricted in their possibilities. When tags follow a command, they usually appear at the end of the IP rather than having their own intonational phrase and the tone can either be a rise or a fall. However, if they do have their own IP, the tone is usually "an encouraging rise, giving a softening effect" (Wells, 2006, p. 50; Cruttenden, 2014, p. 296). Example 6 below, from Wells (2006), illustrates this intonational meaning.

(6) Open the \window, | /would you please?
(Would you please open the /window?)

Moreover, as observed by Wells (ibid), tags with their own IP, which have a falling tone after a command, sound very insistent.

(7) Answer the \phone, | \will you?
(= Will you answer the \phone. || Obey me im\mediately)

Question tags are a very common resource in spontaneous oral speech and are distinctive markers of orality, whose pragmatic and illocutionary function is substantial in oral communication. They can work as triggers for irony and sarcasm, doubt, and confirmation depending on their polarity, and, most importantly, on their tone direction. Consequently, the aim of this paper is to describe how the AI voices from Amazon's TTS system Polly interpret question tags and if they apply the same tone rules as humans do, given the increase in popularity and the professional applications that this software offers.

3. Neural TTS software

Commercial TTS systems are often confined to speaking in predefined voices, based on models created in advance using a time-consuming and non-scalable procedure. Typically, as pointed out by Kons et al. (2018, p. 290), a voice model is constructed from a huge corpus of audio recordings of a single speaker. Using the voice model, the system recreates the recorded speaker's voice. This inflexibility conflicts with the requirement of consumers for TTS to reproduce the voices and speaking styles of their favourite speakers. Such requests may originate from clients seeking distinctive TTS sounds or from the creators of amusing artificial agents speaking in the voices of iconic film heroes.

Emerging neural speech synthesis models seem to offer a potential basis for the development of flexible TTS systems that are readily responsive to the voices of unseen speakers. Amazon Polly's Neural TTS (NTTS) technology is capable, according to the information provided on their website, of producing sounds of even greater quality than its normal voices from their standard TTS. The NTTS technology generates text-to-speech voices that are the most natural and human-like conceivable.

[2] A low rise tone involves a rising pitch movement from a low pitch to a mid pitch (Wells 2006, p. 222).

Standard TTS voices are generated via concatenative synthesis. This approach concatenates the phonemes of recorded speech to generate speech that sounds highly genuine. However, its quality diminishes by the unavoidable changes in speech and the procedures employed to segment the waveforms.

Amazon's Polly Neural TTS technology does not synthesize speech using conventional concatenative synthesis. It consists of two parts: a neural network that transforms a series of phonemes, the most fundamental units of language, into spectrograms, which are snapshots of the energy levels in various frequency bands. It also includes a vocoder that transforms spectrograms into an uninterrupted audio output. A sequence-to-sequence model is the initial component of the neural TTS system. This model does not derive its output merely from the matching input, but additionally it examines how the order of the input pieces interact. Then, the model selects its output spectrograms so that their frequency bands highlight acoustic characteristics that the human brain utilizes to process speech. The model's output is subsequently sent to a neural vocoder, and the spectrograms are converted into voice waveforms.

4. Corpus and methodology

To conduct our descriptive analysis, 40 tag questions were used as input. These were produced by the British and American English voice skins available on Amazon Polly. The 600 audio files produced by these voices were downloaded and analysed using the speech analysis software *Praat* (Boersma & Weenink, 1992–2023), which allowed us to see the pitch contours of the tags and describe them as rise, fall, and *odd*. Although there is no tonal category known as *odd*, this label was employed to refer to some intonation patterns we have found in the analysis which are not used in tag questions by English native speakers. We have noticed that some voice skins delivered certain input lines with a misplaced tonic syllable (falling on the pronoun) or using a flat and levelled intonation. These tone patterns are odd and unusual. Therefore, we decided to label such instances as *odd* in our analysis.

Once the results from the study were obtained, we applied some basic statistics analysis to see the behaviour of the voice skins when faced with tag questions and how they utter them.

The 40 tag questions were divided into three categories: reverse polarity tags, copy tags, and commands. The reverse polarity tags were, in turn, divided into two more categories: positive tags and negative tags, each with 10 examples each. Thus, the corpus includes 10 copy tags, 10 command tags, 10 reverse polarity positive tags, and 10 reverse polarity negative tags. The tag questions used in this work are shown in Table 1.

The voice skins selected for this project are the 4 British voices (Emma, Amy, Arthur, and Brian), and the 11 American English voices (Ivy, Joanna, Kendra, Kimberly, Ruth, Salli, Joey, Justin, Kevin, Mathew and Stephen) available on the software at the time of our study. There are 8 female voices and 7 male voices. Overall, a total of 600 utterances were analysed.

The pitch contour for each utterance was analysed by means of *Praat*, and the software provided us with graphic evidence of the behaviour of each voice skin in terms of tone direction. Table 2 shows the pitch contour for the first sentence of positive tags uttered by the British voices.

As can be seen in Table 2, the pitch contours clearly show that there is one falling intonation (Brian's) and three rising (Amy's, Emma's, and Arthur's). Moreover, from the three rising tags, we can observe that Arthur and Emma use the standard yes-no rise, while Amy delivers the utterance with a fall-rise, which is rather unusual, as it has not been described in the literature concerning tags with their own IP.

Reverse polarity tags	
Positive tag	**Negative tag**
1 I didn't know she was a great cook, did you? 2 I've heard that some people believe the Earth is flat, but that's not possible, is it? 3 Most EFL teachers don't focus on phonetics in their classes, do they? 4 If I fail this test now, I won't get another chance, will I? 5 Mum said we're going to New York for Christmas, but we aren't, are we? 6 After hearing you speaking French, I assume you're not bilingual as you claim on your CV, are you? 7 The bread isn't yesterday's, is it? 8 You and I don't have a lot of things in common, do we? 9 I'm not your type, am I? 10 Snow isn't black, is it?	1 Look, Martin has spilled the coffee again. He's so clumsy, isn't he? 2 Friends is a great TV series, isn't it? 3 Most people like chocolate, don't they? 4 This is the best ice-cream ever, isn't it? 5 It's so sad she had to leave the company, isn't it? 6 Here's my famous spaghetti a la bolognese. I'm a great cook, aren't I? 7 Jigoro Kano invented judo, didn't he? 8 We're running late, aren't we? 9 They're the best example of beautiful football, aren't they? 10 The earth is bigger than the moon, isn't it?
Copy tags	**Command tags**
1 So you're having a baby, are you? That's wonderful! 2 She wants to marry him, does she? Some chance! 3 So you think that's funny, do you? Think again. 4 Take a seat, won't you? 5 Help me, can you? 6 Help me, can't you? 7 Close the door, would you? 8 Do it now, will you? 9 Let's go get a beer, shall we? 10 Hey Shawn, lend me a hand, will you?	1 Take a seat, won't you? 2 Help me, can you? 3 Help me, can't you? 4 Close the door, would you? 5 Do it now, will you? 6 Let's go get a beer, shall we? 7 Hey Shawn, lend me a hand, will you? 8 Come over here a minute, will you? 9 Open the window, would you please? 10 Answer the phone, will you?

Table 1. List of reverse polarity tags, copy tags, and command tags used in the study.

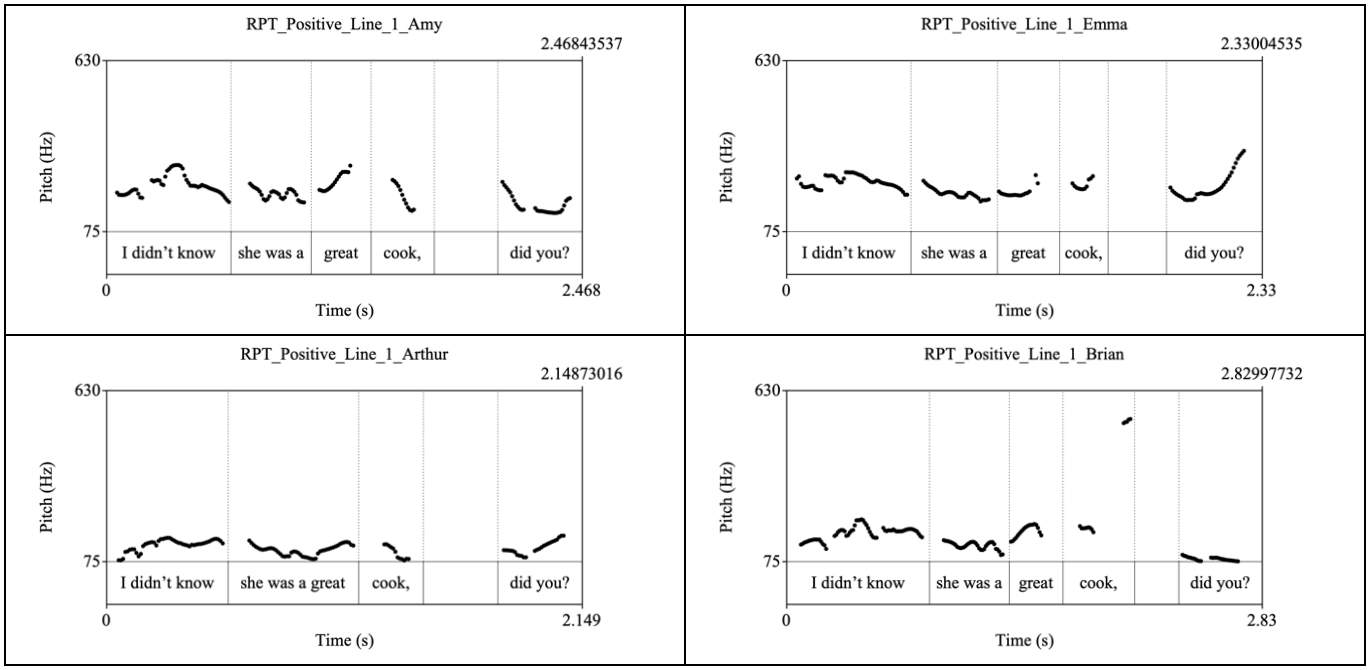

Table 2. Pitch contours for line 1 uttered by the British voice skins Amy, Emma, Arthur, and Brian.

5. Analysis

This section analyses the utterances produced by the British and American English voices. We will comment on each tag question type (reverse polarity, copy tag and command), start with the quantitative analysis to get the overall numbers, and continue with the qualitative analysis which will provide us with information about the pragmatic and illocutionary load these utterances convey in terms of tone direction (rise, fall, and odd). The examples identified having an odd intonation in our study show a pitch contour which is irregular and unexpected in these types of constructions. Mainly, these utterances show a misplacement of the tonic syllable, which instead of falling on the auxiliary verb — with the consequent tail formed by the pronoun — falls on the pronoun, with a fall tone, which has no tail[3]. Moreover, this misplacement of the tonic syllable creates narrow focus in terms of tonicity. In addition, some examples of odd tones include tags with flat and levelled intonation, a tone pattern that has not been found in the consulted literature concerning these constructions.

5.1. Reverse polarity tags

5.1.1. Reverse polarity tags with positive tag

Quantitative results

Out of the 150 utterances produced by Polly, 107 were uttered using a rise, 23 show a falling intonation, and 20 have an odd intonation. This means that 71.3% rise, 15.3% fall, and 13.3% are odd, as shown in Figure 1. The results show almost the same number of falling and odd utterances.

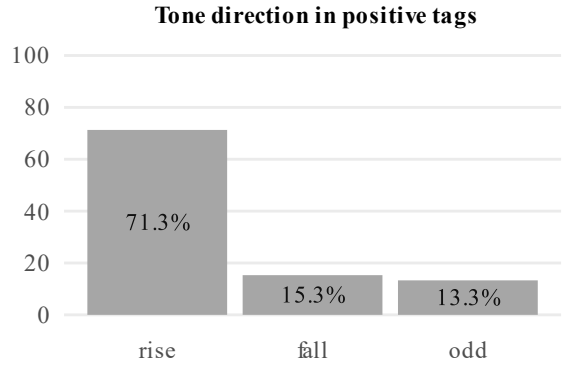

Figure 1. Percentage distribution in reverse polarity positive tags.

Qualitative results

As displayed in Figure 1, almost 72% of the voice samples show a yes-no rise. This indicates that the illocutionary force of these utterances is genuinely asking for information, rather than seeking confirmation or making a statement, as observed in the 16% of the utterances that have a falling tone. From the 107 utterances that are delivered with a rising intonation, 3 of them, uttered by the British voice Amy, show a fall-rise. The pitch contours for these sentences are shown in Table 3.

We can observe that for each utterance the tag has its own IP with a clear fall-rise tone. The tonic syllable falls on the auxiliary verb, where the tone falls, and then it rises along the tail formed by the pronoun. This type of tone pattern is not characteristic of tag questions that have their own IP. According to Wells (2006, p. 49), these patterns can be found in constructions where the anchor and the tag make one single IP and the fall-rise befalls on the word preceding the tag, as in the following example by Wells (ibid.).

(8) So you've qualified as a \/<u>law</u>yer, have you?

[3] The part of an IP that follows the nucleus is known as the *tail* and it contains no accented syllables. If the nucleus is located on the last syllable of the IP, there is no tail (Wells 2006, p. 8).

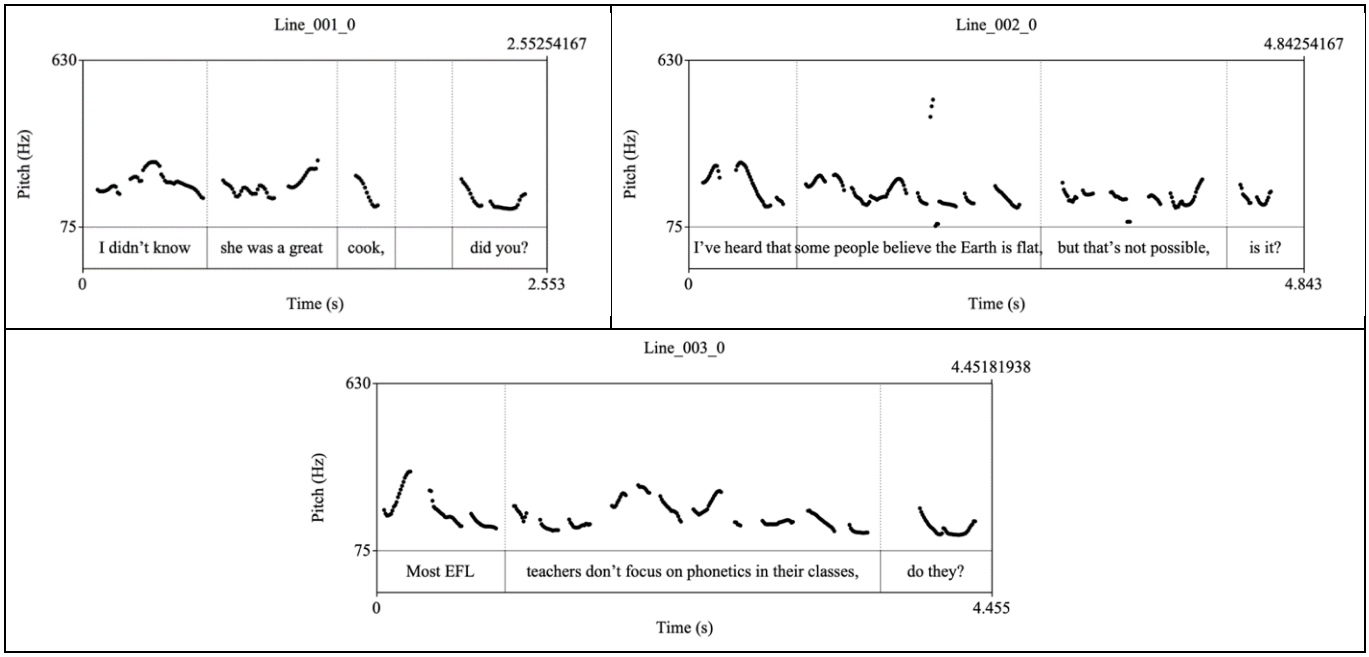

Table 3. Fall-rise pitch contours in British voice skin Amy.

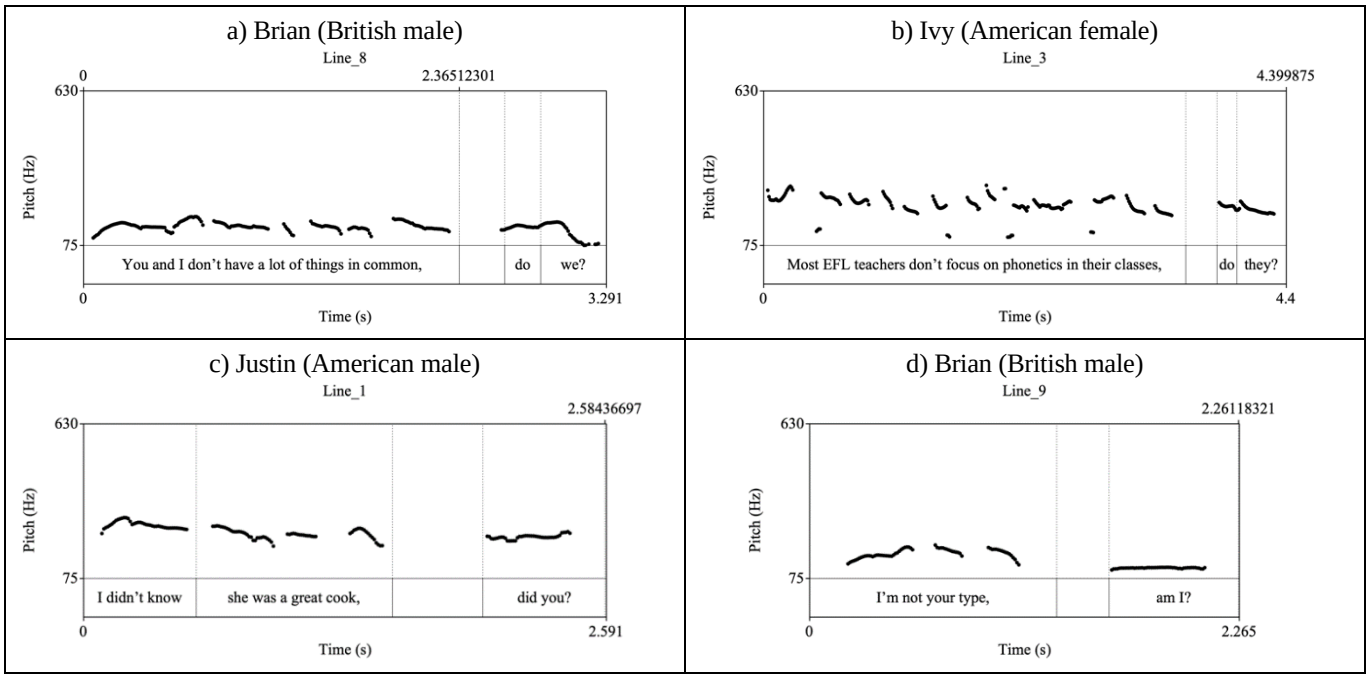

Table 4. Examples of odd intonation by the voice skins Brian, Ivy, and Justin.

Another significant result is that 13% of the utterances show an odd intonation, where the tonic syllable is mostly displaced, as shown in Table 4.

The images shown in Table 4 clearly depict the odd intonation patterns of the tags. While c and d show tones that are completely flat and monotonous, the contours in a and b show a displacement of the tonic syllable, which falls on the pronoun rather than on the auxiliary verb, as would be expected. These utterances, despite having a falling tone on the pronoun, sound odd because the auxiliary is not accented, and there is no tail to continue the melody of the tone that falls on the monosyllabic pronouns.

5.1.2. Reverse polarity tags with negative tag

Quantitative analysis

The analysis for the negative tags shows that out of the 150 utterances delivered by Polly's voices, 95 have a rise, 28 a fall, and 27 an odd intonation. In terms of percentages, this means that 63.3% rise, 18.7% fall and 18% sound odd. As seen in the previous section, there is almost the same number of falling and odd tags, and a high percentage of rises. Figure 2 illustrates these numbers.

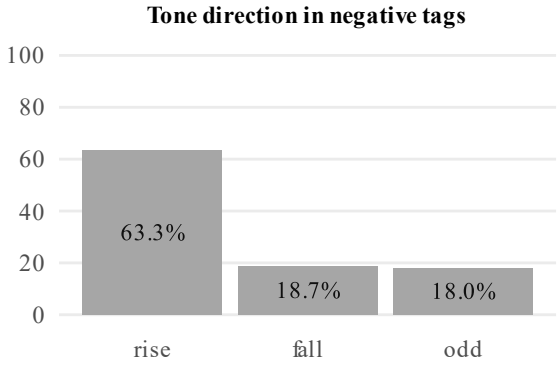

Figure 2. Percentage distribution in reverse polarity negative tags.

Qualitative analysis

Most of the utterances show a yes-no rising intonation, which implies that this type of tag will be frequently uttered requesting an answer from the receiver. Only 18.7% of the times Polly's voices deliver this kind of construction will be requesting confirmation or just making a statement by means of falling intonation. Once again, there is an important number of utterances showing odd intonation, either because the tone is flat and monotonous or because the tonic is displaced, as happened with positive tags. Some examples of the three intonational possibilities for sentence number 5 are displayed in Table 5.

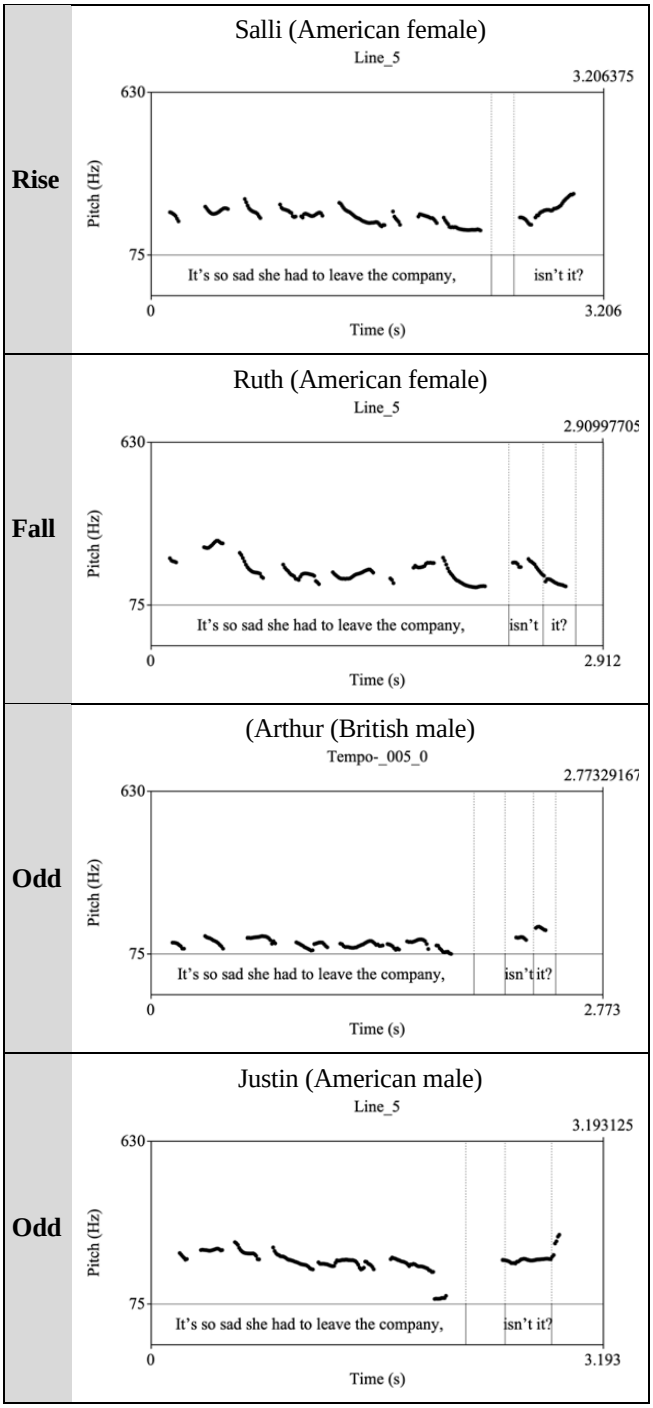

Table 5. Pitch contour samples for reverse polarity with negative tag sentence number 5 (*It's so sad she had to leave the company, isn't it?*).

As displayed above, this line can be uttered using a yes-no rise as shown in Salli's contour, a definite fall as in Ruth's, or an odd tone as in Arthur's and Justin's contours. The odd deliveries depict narrow focus tonicity in the sample uttered by Arthur (the tonic syllable falls on *it*), and a flat monotone delivery in Justin's utterance. There is, however, some

sort of rising tail in Justin's contour. However, this rising movement is a microprosodic effect and belongs to the frequencies of the phoneme /t/ during the plosion.

Line 5 exemplifies the main trend present in the delivery of the 10 sentences used as input. There is no uniformity as to what tone direction will be used by each of the voice skins for each input line. We have noticed that Emma, Joanna, and Mathew use a rising intonation for all their utterances, unlike the other voice skins, that use rising, falling and odd intonation.

5.2. Copy tags

Quantitative analysis

The results for the analysis of the 150 utterances of copy tags show that 110 of these were delivered using a rising intonation, 30 a falling one, and 13 an odd tone. While 73.3% of the input sentences were uttered using a rise, which is the expected intonation pattern when tags have their own IP (Wells, 2006, p. 49), 26.7% do not follow this pattern and are uttered with either a falling tone (20%) or an odd intonation (6.7%). The bar chart in Figure 3 shows the percentages for the copy tags.

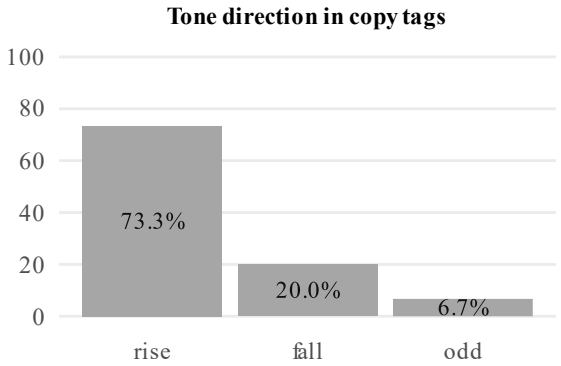

Figure 3. Percentage distribution in copy tags.

Qualitative analysis

The results obtained in the quantitative analysis indicate that when using copy tags in Amazon Polly, there is a 30% probability of having them uttered in an unexpected and unnatural way. All the utterances produced by Brian (British) were delivered either with a fall or an odd tone, which indicates this voice skin is not ready for this kind of constructions yet. Nonetheless, within the 70% of successful deliveries we have noticed that there are four voice skins that managed to utter all the copy tags with a rising tone: Emma (British), Joanna (American), Kevin (American), Matthew (American), and Stephen (American).

Table 6 includes the tone types used by some of the voice skins for sentence 9.

The pitch contours displayed in Table 6 show that all the tags have their own IP and their own tone. We can see that the first two contours (Arthur and Kevin) show rising tones, as expected in these constructions. The next three contours break Cruttenden's (2014) and Wells's (2006, p. 49) rule — "constant-polarity tags, if they have their own tone, always have a rise" — as two falls and one odd are produced. Amy's tag contour shows a high-fall tone, while Joey's is a low-fall one. Kendra's tone sounds odd due to the displacement of the tonic syllable, which falls on the pronoun *they* instead of the auxiliary *did*. Again, we find an instance of narrow focus in tags, as in the analysis of reverse polarity tags.

5.3. Command tags

Quantitative analysis

The figures concerning command tags show that out of the 150 utterances, 121 rise, 22 fall, and 7 have an odd intonation. In terms of percentages, 80.7% rise, 14.7% fall, and 4.7% are odd. Once more, the intonational rule is broken 20% of the time. According to Wells (2006, p. 50), tag questions after commands that do not have a rising intonation may be considered as not being well-formed by some speakers. In addition, this scholar indicates that, when a tag following a command has its own IP, the usual tone is "an encouraging rise, with a softening effect".

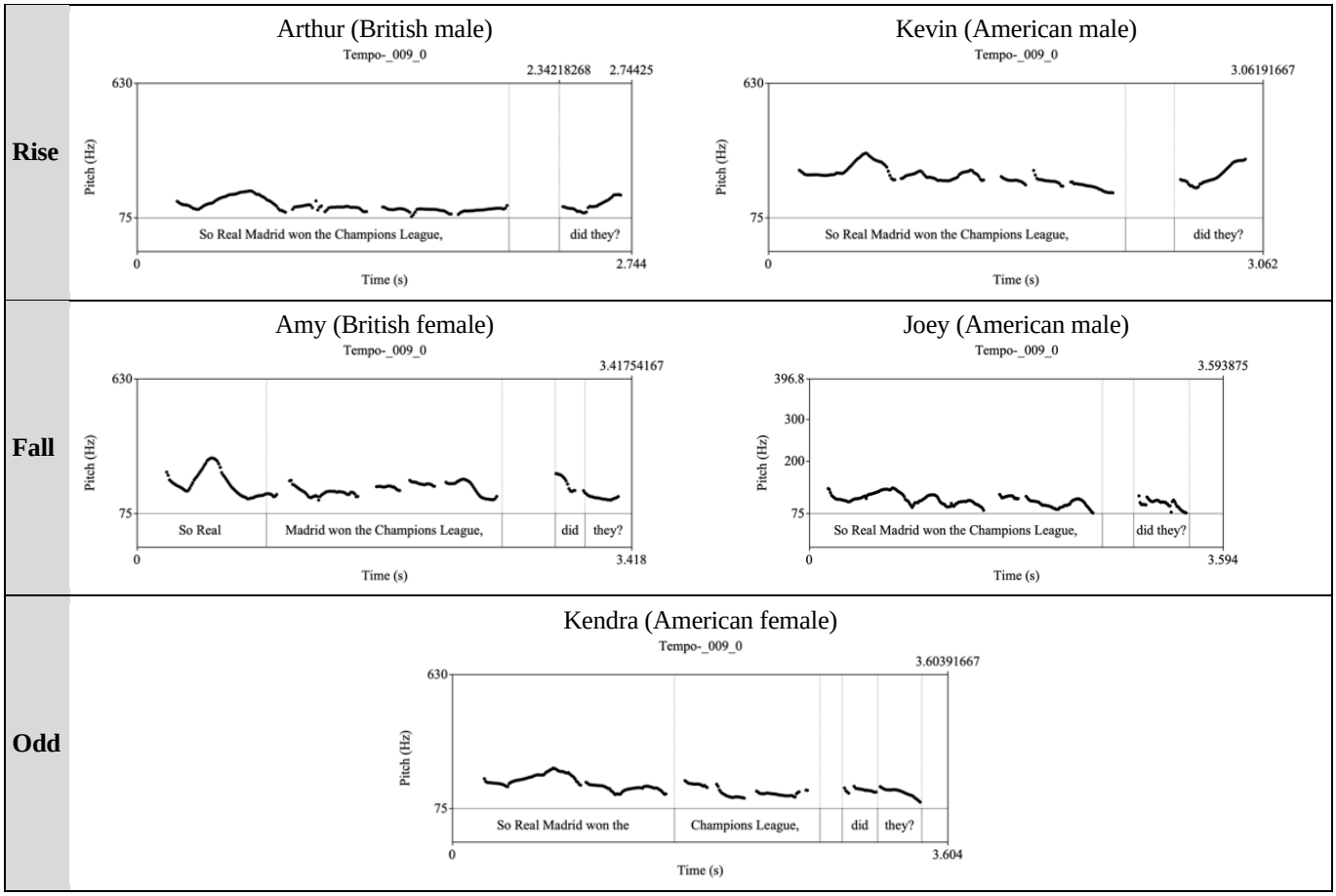

Table 6. Pitch contour samples for copy tag sentence #9 (*So Real Madrid won the Champions League, did they?*).

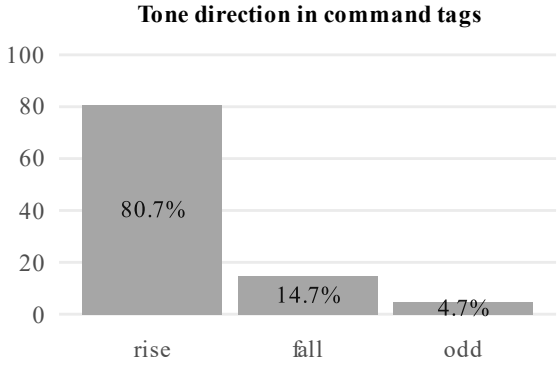

Figure 4. Percentage distribution in command tags.

Qualitative analysis

After analyzing the 150 utterances, we have observed that all of them, except for Amy's and Emma's line 5, have the tag in a separate intonational phrase (IP). These are the only examples of tags without their own IPs in the whole analysis (including the other tag types). This time, the voice skins decided to deliver the whole line as one IP, obviating the comma in the text. Table 7 displays Amy's and Emma's utterance compared with Matthew's and Kimberly's so we can see the difference. We have included the audio waves in the images to have a better representation of the audio delivery.

It can be appreciated from the images on Table 7 that Amy's and Emma's utterances are made up of a single IP, while Kimberly's and Matthew's contain two. The audio waves for the first two utterances (Amy's and Emma's) show no pause, each lasting 1.3 and 1.2 seconds. On the other hand, Kimberly's and Mathew's audio waves show the gap for the pause after the comma, thus splitting the utterance into two IPs.

A detailed analysis of Amy's pitch contour shows that despite the fact that there is one IP and the tag is attached to the anchor, the tonic syllable falls on the tag, on *will*, with a high fall nuclear tone, with *you* as a low tail. Emma's pitch contour exhibits a similar pattern where the nuclear tone falls on the

auxiliary verb *will* in the tag. This falling tone is followed by a rising tone that continues rising along the tail, formed by the pronoun *you*.

Apart from the two examples concerning the tonality differences of the command tags, we have found two instances of narrow focus and tonic displacement in the tags, that called our attention. The voice skins responsible for this kind of delivery are Brian (line 6) and Ruth (line 10). Table 8 shows the pitch contour for their respective utterances.

The pitch contours over the tags in the utterances represented in Table 8 illustrate the narrow focus tonicity of these sentences. Brian's delivery shows a standard fall on *we*, while Ruth's is a clear example of a high fall since the pitch level goes as high as 300 Hz. Apart from not sounding well formed, as Wells (2006, p. 50) considers, because of the falling intonation, the fact that the tonic syllable is displaced thus creating narrow focus, makes the utterances sound even more odd.

In this section we have seen examples of the most relevant utterances for each tag question type. After analysing the results of the research, we can now draw some conclusions concerning the delivery of tag questions and their illocutionary force by Amazon's AI Polly.

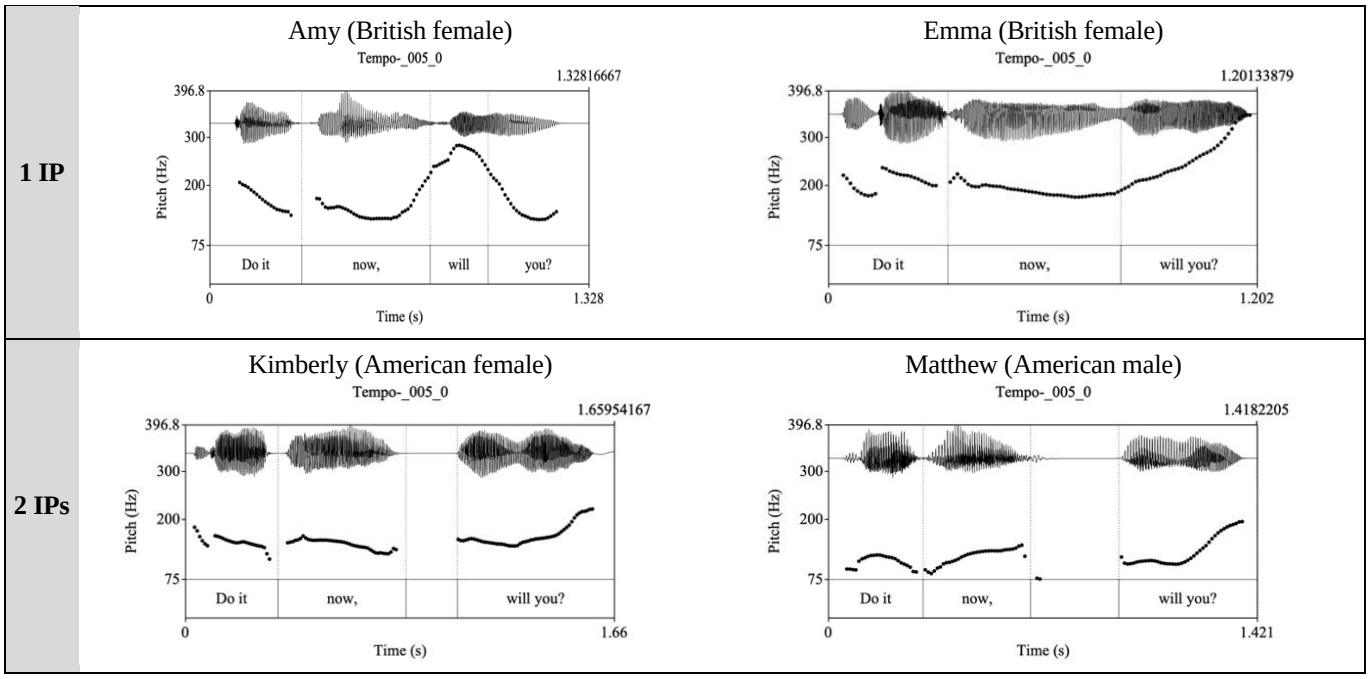

Table 7. Examples of IP structure for command tags in sentence #5 (*Do it now, will you?*).

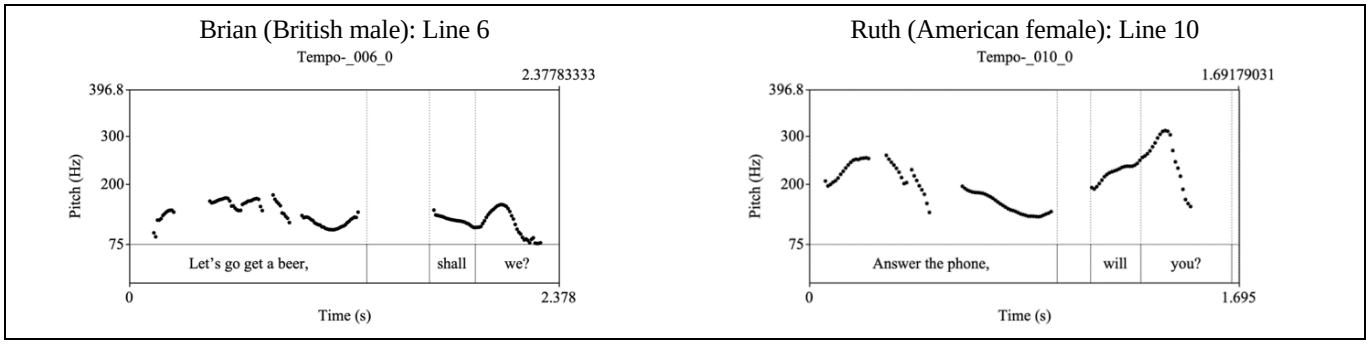

Table 8. Pitch contours of narrow focus command tags.

6. Conclusions

This study has shown that the most recurring tone pattern used by the voice skins from Amazon Polly is the rise. Regarding the whole samples of tag questions produced by the software, there were 600 utterances in total. 433 were produced with a rising intonation, 103 had a falling tone and 64 were delivered with an odd or irregular pronunciation, as illustrated in Figure 5 below.

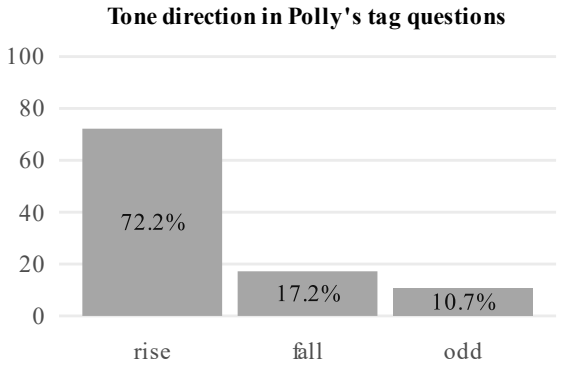

Figure 5. Distribution of tone direction in Polly's tag questions.

One of the conclusions that can be drawn from these results is that the software mainly treats these kinds of constructions as genuine true questions since 72.2% of the utterances have a rising tune. This coincides with Wells (2006), Roach (2009) Parrot (2010), Collins and Mees (2013), Cruttenden (2014), and with other scholars consulted on the pragmatic meaning of the rising intonation for tag questions. It seems as if the software rises the intonation of the utterance automatically when it identifies a question mark. However, question tags do not always have to be uttered with a rising intonation and deliver the illocutionary force of a yes-no question. In fact, as stated by Estebas Vilaplana (2014, p. 278), most question tags (referring to reverse polarity tags) have a falling intonation and work mainly as confirmation requests. Nonetheless, despite an expected falling tone in a reverse polarity, we have seen that the most common tonal pattern for reverse polarity tags in Polly is a rise. The same applies for copy tags and command tags, which are usually uttered with a rise both by humans and the AI software.

If command and copy tags are inputted into this software, there is approximately 75% probability of delivering them with the expected rising intonation. On the other hand, if what we input are reverse polarity tags, the chances that the software delivers an undesired tone are as high as 85%, given that the falling tone appears 15% of the time. This means that Amazon's Polly TTS AI goes against the general trend concerning the tone pattern for these reverse polarity constructions, which implies delivering an illocutionary force that goes (mostly) against the expected one.

Another conclusion from our analysis is that an odd intonation pattern is used in 10.7% of the tags. As we have seen, there are several examples of narrow focus tonicity (the nuclear tone falling on the pronoun rather than on the auxiliary verb), which violates the accentual rule of tags. Collins and Mees (2013) refer to this possible odd intonation pattern as sometimes being produced in error by non-native speakers of English. They consider that "in all tags, the nucleus *invariably* falls on the verb — *never* on the pronoun. An intonation pattern such as the following, with the pronoun as nucleus is completely unacceptable in English […]" (Collins & Mees, 2013, p. 152). This conception is also shared by Gómez González and Sánchez Roura (2016). Consequently, 10.7% of the utterances produced by the software will either have a flat monotone pattern, inexistent in the literature that studies tag questions, or be delivered with a completely unacceptable English pattern. In both cases, the pragmatic load from the tags will be undesirable by the script writer and software user.

In addition to the unusual tone pattern found in our corpus, we must comment on the unexpected falling tone of copy tags and command tags. Most scholars believe that copy tags are typically uttered with a rising intonation. Cruttenden (2014, p. 296) goes even further and believes that for copy tags "falling tones are impossible". Therefore, based on the results obtained in this study, there is a 26.7% chance of delivering an inconceivable tone while using this software, from which 6.7% are completely unacceptable (delivered either with flat tone or narrow

focus) and 20% impossible as they are uttered with a falling intonation.

Concerning command tags, the results show, as in the case of copy tags, that there is 20% chance to obtain an unexpected tone delivered. Wells (2006, p. 50) together with Gómez González and Sánchez Roura (2016), understand that the correct way of uttering this type of tag is by using a rise, which softens the force of the command in positive tags and conveys a more demanding attitude on the part of the speaker in negative ones. In both cases, command tags have the illocutionary force of a request, not a question (Gómez González & Sánchez Roura, 2016, p. 308). Delivering this kind of tag with an unforeseen tone pattern, affects directly the pragmatic meaning of the utterance and will inevitably alter the speaker's intentions.

This work has provided a descriptive account of the intonation used by Amazon Polly's voice skins when faced with different forms of tag questions. Almost all the voice skins seem to be able to provide different tone patterns, which means that they are aware of the different tone possibilities for tag questions. There are some voices, mainly the male American ones, that tend to provide a rising intonation for all the tag types. More research on the delivery of tag questions should be conducted in terms of gender and accent, and even voice type (kids, young, middle age, etc.), and include other types of AI software.

The use of AI powered services like ChatGPT for content creation, DeepL for translations, and TTS such as Amazon's Polly, is here to stay and will shape the future of different sectors like translation and education. However, the possible usage of TTS for learning and entertaining purposes, where spontaneous language is present, seems to be distant if we are to deliver a coherent discourse that conveys the right illocutionary force, which is the ultimate goal in communication. How can the voice skin decide which tone to use when facing tag questions? Does the script writer or software user have to manually modify the software code so that the tone for each tag behaves as it should? We have not been offered the chance to modify the tone direction of any of the tags while using the software and it seems improbable that users will tune each tone for each utterance for a successful delivery of the illocutionary force while working with Polly. Amazon's Alexa developer documentation, available online[4], prides itself on improving the interactivity and customer experience using its NTTS voices. In the light of the results of our work, it may not be advisable for Polly's users to include question tags in their scripts. This limitation restricts the usage of spontaneous natural language. If users wished to have a script recorded with natural speech traits such as question tags for e-learning modules, it would be advisable for them to record it themselves or hire a professional human voice talent who will certainly understand the communicative intent of each tag.

References

Boersma, P., & Weenink, D. (1992–2023). *Praat: Doing phonetics by computer* (Version 6.3.18) [Computer program]. https://www.praat.org/

Cattel, R. (1973). Negative transportation and tag questions. *Language, 49*, 612–639. https://doi.org/10.2307/412354

Cohen, M. H., Giangola, J. P. & Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.

Collins, B., & Mees, I. M. (2013). *Practical phonetics and phonology. A Resource Book for Students*. Routledge. https://doi.org/10.4324/9780203080023

Cruttenden, A. (2014). *Gimson's Pronunciation of English*. Routledge. https://doi.org/10.4324/9780203784969

[4] Alexa Developer Documentation: https://developer.amazon.com/en-US/docs/alexa/custom-skills/guidelines-ux-amazon-polly-skills.html

Estebas Vilaplana, E. (2014). *Teach yourself English pronunciation: An interactive course for Spanish speakers*. Universidad Nacional de Educación a Distancia.

Gómez González, M. A., & Sánchez Roura, M. T. (2016). *English pronunciation for speakers of Spanish: from theory to practice*. Walter de Gruyter. https://doi.org/10.1515/9781501510977

Kay, P. (2006). Pragmatic aspects of grammatical constructions. In L. R. Horn, & G. Ward (Eds.), *The Handbook of Pragmatics*. (pp. 675–700). Blackwell Publishing.

Kim, H., Kim, S., & Yoon, S. (2022). Guided-TTS: A diffusion model for text-to-speech via classifier guidance. *Proceedings of Machine Learning Research, 162* [*Proceedings of the 39th International Conference on Machine Learning*], 11119–11133.

Kons, Z., Shechtman, S., Sorin, A., Hoory, R., Rabinovitz, C., & da Silva Morais, E. (2018). Neural TTS voice conversion. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 290–296). IEEE. https://doi.org/10.1109/SLT.2018.8639550

Lakoff, R. (1969). A syntactic argument for negative transportation. In R. I. Binnick, A. Davidson, G. M. Green, & J. L. Morgan (Eds.), *Papers from the 5th Regional Meeting of the Chicago Linguistic Society* (pp. 140–147). Department of Linguistics, University of Chicago.

Leech, G., & Svartvik, J. (1994). *A communicative grammar of English*. Longman. https://doi.org/10.4324/9781315836041

Mateo, M. (2014). Exploring pragmatics and phonetics for successful translation. *VIAL (Vigo International Journal of Applied Linguistics), 11*, 111–135.

McCawley, J. D. (1988). *The syntactic phenomena of English*. University of Chicago Press.

Mott, B. (2011). *English phonetics and phonology for Spanish speakers*. Publicacions i Edicions de la Universitat de Barcelona.

Parrot, M. (2010). *Grammar for English language teachers*. Cambridge University Press. https://doi.org/10.1017/9781009406536

Roach, P. (2009). *English phonetics and phonology: A practical course*. Cambridge University Press.

Rodríguez Fernández-Peña, A. C. (2022). La equivalencia pragmática de las 3Ts en inglés y español. *LynX: Panorámica de estudios lingüísticos, Extra 25* [*Gramática Contrastiva: Métodos y Perspectivas*, ed. M. A. Lledó], 177–218.

Sadock, J. M. (1974). *Toward a linguistic theory of speech acts*. Academic Press.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., & Skerrb-Ryan, R. (2018). Natural TTS Synthesis by conditioning WaveNet on MEL spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779–4783). IEEE. https://doi.org/10.1109/ICASSP.2018.8461368

Swan, M. (2005). *Practical English usage*. Oxford University Press.

Tench, P. (2009). *The pronunciation of grammar* [Conference presentation]. 3rd International Congress on English Grammar. Salem, TN, India.

Thomson, A. J., & Martinet, A. V. (1986). *A practical English grammar*. Oxford University Press.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30 (NIPS 2017)* (pp. 6306–6315). Curran Associates Inc.

Vince, M., & Emmerson, P. (2003). *First Certificate language practice*. Macmillan Education.

Wells, J. C. (2006). *English intonation. an introduction*. Cambridge University Press.