Representing 'how you say' with 'what you say': English corpus of focused speech and text reflecting corresponding implications

Naoaki Suzuki, Satoshi Nakamura

Nara Institute of Science and Technology, Japan

{suzuki.naoaki.sg4, s-nakamura}@is.naist.jp

Abstract

In speech communication, how something is said (paralinguistic information) is as crucial as what is said (linguistic information). As a type of paralinguistic information, English speech uses sentence stress, the heaviest prominence within a sentence, to convey emphasis. While different placements of sentence stress communicate different emphatic implications, current speech translation systems return the same translations if the utterances are linguistically identical, losing paralinguistic information. Concentrating on focus, a type of emphasis, we propose mapping paralinguistic information into the linguistic domain within the source language using lexical and grammatical devices. This method enables us to translate the paraphrased text representations instead of the transcription of the original speech and obtain translations that preserve paralinguistic information. As a first step, we present the collection of an English corpus containing speech that differed in the placement of focus along with the corresponding text, which was designed to reflect the implied meaning of the speech. Also, analyses of our corpus demonstrated that mapping of focus from the paralinguistic domain into the linguistic domain involved various lexical and grammatical methods. The data and insights from our analysis will further advance research into paralinguistic translation. The corpus will be published via LDC and our website [1].

Index Terms: English, paralinguistic information, focus in speech and text, corpus, speech translation

1. Introduction

In speech communication people make use of two types of information to convey their intentions: *what* is said (linguistic information) and *how* it is said (paralinguistic information) [1]. Paralinguistic information is expressed by suprasegmental features such as duration, intensity and pitch. Even with the same linguistic information, changes in these prosodic features can communicate different implications. One such implication involves emphasis. As the terminology can be ambiguous, we follow Kohler's [2] distinctions of two kinds of emphasis: *emphasis for focus*; which 'singles out elements of discourse by making them more salient than others'; and *emphasis for intensity*; which 'intensifies the meaning contained in the elements.' The current work addresses the first category of emphasis: focus. In the literature of a semantic framework called Alternative Semantics [3, 4], focus is understood as the indication of 'the presence of alternatives that are relevant for the interpretation of linguistic expressions' [5]. Consider this example:

(1) a. **John** bought the apple.

  b. John **bought** the apple.

In (1a), *John* is focused, as indicated by bold typeface. The speaker implies that *It was John who bought the apple*, indicating the presence of contextually possible alternatives such as *Peter* or *Mary*, and at the same time indirectly ruling out these agents for the person who *bought the apple*. In contrast, in (1b), focus falls on *bought*, by which the speaker indicates that *What John did was not sell the apple, but buy the apple*. Although (1a) and (1b) give identical linguistic information, the differences in the focused words create different connotations. In English speech, focus is marked by sentence stress, the most prominent stress within a sentence [6]. To avoid misunderstanding, interlocutors must correctly perceive the placement of sentence stress and understand the associated focused meaning. While this inherent skill is taken for granted among native English speakers, non-natives find it challenging [7]. Therefore, cross-lingual interactions demand an automated system that can output the implied meaning for non-natives.

The idea of speech translation (ST), which automatically translates speech in a source language (SL) to text or speech in a target language (TL), might be helpful toward achieving such cross-lingual communication. In recent years, ST systems have made significant progress in translating linguistic information correctly. However, most of the ST systems developed so far have not attained the capability to consider paralinguistic information, including focus. Translation by these systems is based on the transcription produced by automatic speech recognition (ASR), which is designed to transcribe only the contents of an utterance, resulting in the loss of paralinguistic information. The current ST models translate (1a) and (1b) the same way, even though they convey different implications. Recently, however, as the importance of paralinguistic information has become more widely acknowledged, some studies have tackled the translation of paralinguistic information such as voice quality [8], emotions [9, 10] and emphasis [9, 11, 12, 13, 14, 15, 16, 17] by mapping the prosodic cues in the SL, such as intensity, duration and fundamental frequency, to speech in the TL.

While such efforts have advanced the research on translation of paralinguistic cues, it cannot be assumed that the TL necessarily has a prosodic counterpart that plays the same role as that in the SL. For instance, for information focusing, languages make use of not only prosodic devices but also lexical and grammatical devices [18]. Although English speech often uses sentence stress rather than linguistic devices for focus [18], the degree of reliance on such methods vary from language to language [18, 19, 20]. This observation partially limits the efforts of an acoustic-to-acoustic way of paralinguistic translation. Instead, we argue the potential of the acoustic-to-linguistic mapping of paralinguistic information by paraphrasing the speech into the linguistic domain within the SL, with the help of lexical and grammatical devices, and then passing the paraphrases on to the translation module (Figure 1). Such
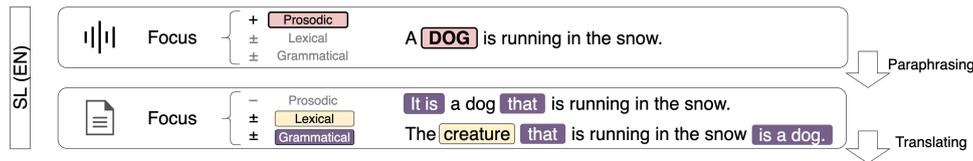

Figure 1: *Example of focus translation through paraphrased text. Before translating, within the SL, focus expressed by prosodic device in the original speech is transformed into that expressed with lexical and grammatical devices in text.*

a method could return different translations depending on the focused items, even with the same linguistic information.

The achievement of acoustic-to-linguistic focus transformation requires data for training such a model. However, to the best of our knowledge, no existing English corpus contains pairs of: speech having different items in focus; and text reflecting the relevant implications. In the most related work, an English corpus [21] was built with text items that differ in the degree of *emphasis for intensity*, e.g. (*It is a little bit hot* / *It is extremely hot*), and the corresponding speech, e.g. (*It is **hot*** / *It is **HOT***) which was recorded so that it would represent the same degree of emphasis as that in the text. Since their focus was on the degree of emphasis such as weak/strong, every speech sample fixed its position of focus on an adjective. The data was later used for building a model of speech-to-text emphasis translation [22]. A recent work in text-to-speech (TTS) [23] added the elements of *emphasis for focus* to the previous efforts. Speech items which differed in the placement of focus were collected, e.g. (***Sarah*** *closed the door*, *Sarah **closed** the door*). However, since the purpose of the work was to achieve prosodic prominence in a TTS model from annotated text, the data did not involve the corresponding implications of speech. Also, while words in a closed-class, e.g. *the, is*, can be focused [24], such cases remain to be addressed.

We also argue the importance of understanding the relationships of focus representations in speech and text, i.e. how focus expressed in speech is paraphrased or transformed into the linguistic domain. In the literature, some studies addressed patterns of general paraphrase alternations, i.e. how an original text item is paraphrased into another one which has approximately the same meanings [25, 26, 27], and others examined how linguistic items can convey focus [28, 3]. However, it has not been clear the methods of mapping focus from the paralinguistic dimension to the linguistic one. To these ends, we present the creation of an English corpus containing speech that differs in the placement of focus, where words in a closed-class are also the targets of focus, and text reflecting the relevant implications. We also perform quantitative and qualitative analyses of the transformation of focus from the paralinguistic domain to the linguistic domain.

1.1. Focus in English

This section briefly summarises how the English language employs prosodic and linguistic devices to convey focus.

1.1.1. Prosodic device

As mentioned in the previous section, English speech uses sentence stress to convey focus. Native English speakers highlight certain information and draw a listener's attention using the following procedure [6]. Depending on the intention, they first break spoken materials into smaller chunks called intonation phrases (IPs). Then, in each IP they select the most important word and put sentence stress on that word's stressed syllable, i.e. syllable that has lexical stress.

1.1.2. Lexical and grammatical devices

Prosody is not the only device to convey focus. Lexical and grammatical devices are also available [18]. As one type of the lexical devices, a group of words called focus particles can convey focus [29]. For instance, *only*, *even*, and *alone* can serve this purpose [3]. Consider the following example:

(2) a: John bought an apple.

   b: John bought only an <u>apple</u>.

Compared to (2a), (2b) explicitly states what *John bought* was neither *an orange* nor *a banana*, resulting in *an apple* being highlighted, as indicated by the underline. Similarly, *let alone* [30], reflexive pronouns such as *himself/herself* [31], *particularly, mainly*, and many more such words [29] can be used as lexical items for focus.

A grammatical device can also perform focus by changing the structure of a sentence. Grammatical items include constructions of cleft (*It was <u>Simon</u> who kicked the door.*), pseudo-cleft (*What Mary bought was an <u>apple</u>*), inversion (*And then appears a <u>bear</u>*), and passivization (*I was bit by a <u>dog</u>*) [24]. It should be noted that grammatical reconstructions often involve shifting the target of focus toward the end of the sentence, since English information structure is governed mainly by the principles *Given-Before-New* and *End Weight* [32].

2. Corpus construction

2.1. Text design

We started the text construction from Flickr8k [33], which consists of over 8000 images that depict actions relating to people or animals. Five text descriptions are given for each image, with over 40,000 annotations. To select optimal captions, we removed sentences that fell under one of the following conditions: including punctuation other than a period; corrected by GECToR [34], a grammatical error correction model; a noun phrase; more than six words; identical to another caption. We left only one caption if an image had multiple captions. Regarding the sentence length, for simplicity, we wanted each speech to have only one sentence stress, meaning that an entire sentence is treated as a single IP. To determine an optimal sentence length for a single IP, we examined the London-Lund corpus of spoken English [35], which consists of half a million words and prosodic transcriptions, including tone units, or IPs. We calculated word length for each IP, and chose 6, which equaled the third quartile in the corpus, as the maximum sentence length. After the filtering, we had 1375 sentences and selected the first 196 captions as the source sentences for our corpus. The text items included *A biker enjoys a coffee*, for instance (Figure 2).
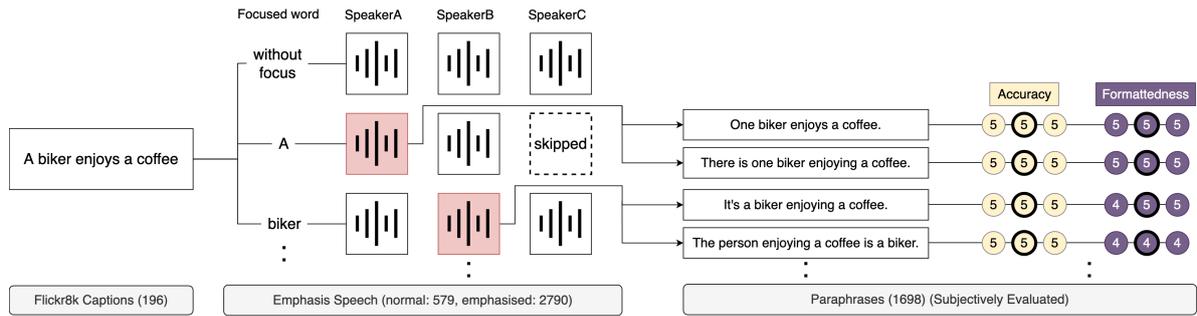
Figure 2: *Corpus Design*

2.2. Speech collection

We employed Amazon's Mechanical Turk (MTurk) for data collection, a crowdsourcing platform allowing researchers to create tasks called HITs (Human Intelligence Task) and anonymous users (Workers) to complete them for a small monetary fee.

2.2.1. Recording application

MTurk does not provide an interface for audio collection, so we created a web-based recording application that could capture a user's recordings interactively and store them in a back-end server. To encourage subjects to speak properly, we enabled Google speech-to-text API so that the system could give the instant feedback 'Speak Clearly' if it did not recognise incoming speech as English words. Speakers could use basic functions such as start and stop of the recordings, and re-recording. As an additional feature, for each utterance, we recorded an extra two seconds at the end while the speech evaluation was in progress; this was done to capture the environment's sound, and during this time, users were instructed to remain silent.

2.2.2. Recording procedure

We attempted to collect a set of speech samples, each with different placement of focus, for each caption using three different speakers (Figure 2). For future use, normal readings without focus was also collected from the same speakers. We asked Workers located in the UK to participate in the tasks. In the recording HITs, Workers were instructed as follows: Look at the displayed written sentence, e.g. (*Two men are ice fishing*); Make a recording emphasising the underlined word if it does not sound unnatural, otherwise, skip this recording; If there is no underlined word, e.g. (*Two men are ice fishing*), read the sentence in a normal way.

When using a crowdsourcing service, taking measures to ensure the quality of data is crucial [33, 36]. We prepared a qualification test, in which Workers needed to fill out general information such as age, gender and accent after an agreement. Then they conducted the step described above for three captions. After manually checking the results, we allowed those who made recordings as instructed to proceed to the main recording HITs. The collection resulted in 3369 speech items (normal: 579, emphasised: 2790) by 10 British natives (5 male, 5 female, age range 22–60 years with a median of 41 years). We paid Workers $1.20 for processing two captions, and $2 for three, resulting in $30.9 per hour on average, surpassing $15, which is considered to be fair among Workers in MTurk [37].
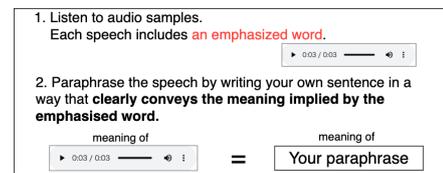
2.3. Paraphrase collection

Figure 3: *Instructions for paraphrase collection*

Free-form tasks, including the paraphrase collection, need to make sure that participants are native speakers [33]. To select native English speakers, we published tasks for Workers residing in an English speaking country (US, CA, UK, AU, NZ) and asked them what their first language was. 600 Workers participated in the task; 550 answered English as their first language. We regarded the 550 Workers as native English speakers, and divide them in half into two groups: G1 and G2.

Figure 3 shows a screenshot of the main instructions given for the paraphrase HITs. Workers were asked to listen to a set of focused speech samples for one caption, where each speech item was randomly selected from available recordings (Figure 2), and then told to come up with a written sentence for that speech item which clearly conveyed the implied meaning in the emphasised speech. We encouraged Workers to freely make changes in grammar and vocabulary but not to use capitalisation, an exclamation mark, or the addition of their own inferences or text providing situational context. We prepared a qualification test for the native speakers in G1, in which they needed to make a brief summary of the instructions in their own words as well as complete the paraphrasing tasks. After manually checking the results, we allowed those who completed the task as instructed to start the main paraphrase HITs. Two workers performed each HIT, obtaining two paraphrases for each emphasised speech sample. We paid $0.15 for each paraphrasing, resulting in $16.5 per hour on average. We obtained 2130 paraphrases from 16 Workers.

2.4. Filtering paraphrases

The quality of the paraphrases produced by the Workers will influence future research using the corpus. To evaluate the quality, we set subjective evaluation HITs. Workers were presented with emphasised speech and a paraphrase and asked to evaluate its quality from two perspectives on a scale from one to five: (A) how accurately the paraphrase conveyed the implied meaning in speech; (B) how well-formatted the paraphrase sentence is. We set a qualification task for the native speakers in G2 in the same

way as the paraphrase collection and asked those who passed to join the main evaluation HITs (16 Workers). Three different Workers evaluated each paraphrase. We calculated median values for both perspectives (Figure 2), and samples with a value of 3 or less for perspective (A) and 2 or less for (B) were discarded. We collected the paraphrases again for those removed samples, and repeated the same procedure of evaluation and removing as above. We paid $0.10 for evaluating each paraphrase, resulting in $25.0 per hour on average.

After the evaluation in MTurk, we filtered out paraphrases which fell one of the following conditions: paraphrases with relatively high variance of the accuracy evaluation scores ($\sigma^2 >$ 1.5), e.g. scores: 2, 5, 5); The tense changed from present to past, e.g. (*The men are climbing.* / *The men were climbing on something.*) After the filtering, we had 1698 paraphrases. Table 1 shows an example of speech and paraphrase pairs.

Table 1: *Example speech-paraphrase pairs in the corpus*

Focused speech	Paraphrase
A biker enjoys a coffee	One biker enjoys a coffee
A biker enjoys a coffee	There is one biker enjoying a coffee
A **biker** enjoys a coffee	It's a biker enjoying a coffee
A **biker** enjoys a coffee	The person enjoying a coffee is a biker
A biker **enjoys** a coffee	A biker drinking a coffee is enjoying it
A biker **enjoys** a coffee	The biker seems to enjoy a coffee
A biker enjoys **a** coffee	A biker enjoys one coffee
A biker enjoys **a** coffee	A biker has one coffee he enjoys
A biker enjoys a **coffee**	What the biker enjoys is a coffee
A biker enjoys a **coffee**	It is coffee the biker enjoys

3. Analysis

To explore how focus in speech was mapped into the linguistic domain, we hand-examined the samples of speech-paraphrase pairs and made broad categorisations of the transformation patterns as follows; from lexical and grammatical perspectives (the patterns do not cover all the transformation methods).

- Lexical transformations

    – **Substitution**: substitute the focused word with its synonyms, e.g. (***dog*** / *canine*), (***on*** / *on top of*).

    – **Modification**: modify the focused word or its phrase with modifiers such as adverbs or a clause, e.g. (***play*** / *play actively*), (***is*** / *is indeed*).

    – **Negation**: explicitly state an alternative of the focused word and negate it (*A **man*** / *A man, not a woman*).

- Grammatical transformations

    – **Leftward shift**: move the focused word towards the beginning of the sentence. Grammatical constructions such as cleft, reversed-pseudo-cleft and inversion were used, e.g. (***People*** *sit ..* / *It's people who sit ..*), (*.. play **baseball*** / *baseball is what .. play*).

    – **Rightward shift**: move the focused word towards the end of the sentence. Grammatical constructions such as pseudo-cleft, inversion and other methods were used, e.g. (*Children **play** ..* / *what children do .. is play*), (*.. is **rock** climbing* / *.. is climbing a rock*)

    – **Tense change**: change the tense from simple to progressive or vice versa, e.g. (***is** walking* / *walks*).

Furthermore, we observed that a certain part-of-speech is more likely to use a certain transformation method. To quantify this tendency, following a method taken for analysis of

paraphrase alternations [26], we randomly sampled 50 paraphrases for each part-of-speech if it had more than 50 occurrences in the corpus and compared the frequency of each transformation method. We counted each phenomenon for each sentence. We restricted the counting only if the focused word or the phrase of the focused word undergoes transformations listed above. Consider this example: (*a dog trots through **the** grass* / *the only grassy area is what a dog trots through*). In this case, not only the focused word *the*, but also the noun phrase it involved, indicated by underline, was under investigation; *only* was inserted (modification), and *grass* was replaced with *grassy area* (substitution). As a grammatical method, reversed-pseudo-cleft (leftward shift) was used. Table 2 shows the average number of occurrences of each transformation in each focused part-of-speech. The pattern of the tense change was counted only when the focused word was verbs or auxiliary verbs.

Table 2: *Mean occurrences of each transformation method per part-of-speech (N: Noun, V: Verb, Adj: Adjective, Num: Numeral, Aux: Auxiliary, P: Preposition, Det: Determiner)*

		N	V	Adj	Num	Aux	P	Det
Lexical	Substitution	0.28	0.12	0.10	0.18	0.08	0.42	0.72
	Modification	0.00	0.02	0.12	0.06	0.60	0.18	0.50
	Negation	0.10	0.06	0.06	0.12	0.10	0.14	0.00
Grammatical	Leftward	0.10	0.20	0.04	0.08	0.00	0.08	0.00
	Rightward	0.46	0.44	0.70	0.52	0.08	0.26	0.02
	Tense	-	0.28	-	-	0.20	-	-

4. Discussion and conclusion

This study set out to create a corpus containing focused speech, where each speech item differs in the placement of focus, and the corresponding text which paraphrases the speech together with its paralinguistically expressed implications. Through our data analysis, we show that the transformation of focus information from the paralinguistic domain to the linguistic domain makes use of a variety of lexical and grammatical devices, and reliance on these methods vary depending on the syntactic category of the focused word. Many of the transformation methods observed in data were the types that have been reported as such devices in the literature, e.g. cleft-constructions in the leftward shift and pseudo-clefting in the leftward shift and negation, which serves the exact purpose of focus defined earlier; indicating the presence of alternatives. On the other hand, we also found some interesting methods such as lexical substitution, e.g. (***man*** / *male adult*).

One of the limitations of the current study is lack of context; *Why do I/they emphasise this word?* In regular paraphrasing tasks, paraphrases and their evaluations can vary depending on context [38]. The importance of context would also be the case for the collection of focused speech and the corresponding text. In future work, we will consider presenting contextual information when collecting recordings and paraphrases.

Despite the limitations, the current work added a new direction toward further improvement of paralinguistic translation; we demonstrated the possibility of mapping paralinguistic information into the linguistic domain. The corpus and insights from our analysis will lead us to construct a ST model which uses the paraphrased text to preserve paralinguistic information.

5. Acknowledgement

6. References

[1] V. Mitra, S. Booker, E. Marchi, D. S. Farrar, U. D. Peitz, B. Cheng, E. Teves, A. Mehta, and D. Naik, "Leveraging acoustic cues and paralinguistic embeddings to detect expression from voice," *arXiv preprint arXiv:1907.00112*, 2019.

[2] K. J. Kohler, "What is emphasis and how is it coded," in *Proc. of Speech Prosody*, 2006, pp. 748–751.

[3] M. E. Rooth, "Association with focus," Ph.D. dissertation, University of Massachusetts Amherst, 1985.

[4] M. Rooth, "A theory of focus interpretation," *Natural Language Semantics*, vol. 1, no. 1, pp. 75–116, 1992.

[5] M. Krifka, "Basic notions of information structure," *Acta Linguistica Hungarica*, vol. 55, no. 3-4, pp. 243–276, 2008.

[6] J. C. Wells, *English intonation: An introduction.* Cambridge University Press, 2006.

[7] T. M. Derwing, R. I. Thomson, J. A. Foote, and M. J. Munro, "A longitudinal study of listening perception in adult learners of English: Implications for teachers," *Canadian Modern Language Review*, vol. 68, no. 3, pp. 247–266, 2012.

[8] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[9] P. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 2006, pp. 557–560.

[10] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Emotional speech recognition and synthesis in multiple languages toward affective speech-to-speech translation system," in *Proc. 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing.* IEEE, 2014, pp. 574–577.

[11] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "A method for translation of paralinguistic information," in *Proc. International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[12] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Generalizing continuous-space translation of paralinguistic information." in *Proc. INTERSPEECH*, vol. 445, 2013, pp. 25–29.

[13] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 153–158.

[14] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "Toward transfer of acoustic cues of emphasis across languages." in *Proc. INTERSPEECH*, 2013, pp. 3483–3486.

[15] Q. T. Do, S. Sakti, G. Neubig, and S. Nakamura, "Transferring emphasis in speech translation using hard-attentional neural network models." in *Proc. INTERSPEECH*, 2016, pp. 2533–2537.

[16] Q. T. Do, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 544–556, 2016.

[17] Q. T. Do, S. Sakti, and S. Nakamura, "Sequence-to-sequence models for emphasis speech translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1873–1883, 2018.

[18] A. Cruttenden, *Intonation.* Cambridge University Press, 1997.

[19] L. J. Downing, "The prosody and syntax of focus in chitumbuka," *ZAS Papers in Linguistics*, vol. 43, pp. 55–79, 2006.

[20] K. Hartmann, "Focus and tone," *Acta Linguistica Hungarica*, vol. 55, no. 3-4, pp. 415–426, 2008.

[21] Q. T. Do, S. Sakti, and S. Nakamura, "Toward multi-features emphasis speech translation: Assessment of human emphasis production and perception with speech and text clues," in *Proc. 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 700–706.

[22] H. Tokuyama, S. Sakti, K. Sudoh, and S. Nakamura, "Transcribing paralinguistic acoustic cues to target language text in transformer-based speech-to-text translation," *Proc. Interspeech 2021*, pp. 2262–2266, 2021.

[23] S. Latif, I. Kim, I. Calapodescu, and L. Besacier, "Controlling prosody in end-to-end TTS: A case study on contrastive focus generation," in *Proc. 25th Conference on Computational Natural Language Learning (CoNLL)*, 2021, pp. 544–551.

[24] S. Greenbaum, *A student's grammar of the English language.* Pearson Education India, 1990.

[25] C. Boonthum, "iSTART: Paraphrase recognition," in *Proc. ACL Student Research Workshop*, 2004, pp. 31–36.

[26] W. Dolan, C. Quirk, C. Brockett, and B. Dolan, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *Proc. 20th International Conference on Computational Linguistics (ACL)*, 2004.

[27] M. Vila, M. A. Martí, H. Rodríguez *et al.*, "Is this a paraphrase? What kind? Paraphrase boundaries and typology," *Open Journal of Modern Linguistics*, vol. 4, no. 01, p. 205, 2014.

[28] J. Taglicht, *Message and emphasis: On focus and scope in English.* Addison-Wesley Longman Limited, 1984, no. 15.

[29] E. König, *The meaning of focus particles: A comparative perspective.* Routledge, 2002.

[30] C. J. Fillmore, P. Kay, and M. C. O'connor, "Regularity and idiomaticity in grammatical constructions: The case of let alone," *Language*, pp. 501–538, 1988.

[31] E. König and V. Gast, "Focused assertion of identity: A typology of intensifiers," *Linguistic Typology*, vol. 10, pp. 223–76, 2006.

[32] B. Aarts, *Oxford modern English grammar.* Oxford University Press, 2011.

[33] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

[34] K. Omelianchuk, V. Atrasevych, A. Chernodub, and O. Skurzhanskyi, "GECToR–grammatical error correction: Tag, not rewrite," in *15th Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 163–170.

[35] J. Svartvik, *The London–Lund corpus of spoken English: Description and research.* Lund University Press, 1990, vol. 82.

[36] R. Kennedy, S. Clifford, T. Burleigh, P. D. Waggoner, R. Jewell, and N. J. Winter, "The shape of and solutions to the MTurk quality crisis," *Political Science Research and Methods*, vol. 8, no. 4, pp. 614–629, 2020.

[37] M. E. Whiting, G. Hugh, and M. S. Bernstein, "Fair work: Crowd work minimum wage with one line of code," in *Proc. AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 197–206.

[38] R. Barzilay and K. McKeown, "Extracting paraphrases from a parallel corpus," in *Proc. 39th annual meeting of the Association for Computational Linguistics*, 2001, pp. 50–57.