User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning

Máté Ákos Tündik[1], *György Szaszák*[1], *Gábor Gosztolya*[2], *András Beke*[3]

[1]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
[2]Research Group on Artificial Intelligence, Hungarian Academy of Sciences, Szeged, Hungary
[3]Spicy Analytics Ltd., Budapest, Hungary
{tundik,szaszak}@tmit.bme.hu

Abstract

Punctuation of ASR-produced transcripts has received increasing attention in the recent years; RNN-based sequence modelling solutions which exploit textual and/or acoustic features show encouraging performance. Switching the focus from the technical side, qualifying and quantifying the benefits of such punctuation from end-user perspective have not been performed yet exhaustively. The ambition of the current paper is to explore to what extent automatic punctuation can improve human readability and understandability. The paper presents a user-centric evaluation of a real-time closed captioning system enhanced by a lightweight RNN-based punctuation module. Subjective tests involve both normal hearing and deaf or hard-of-hearing (DHH) subjects. Results confirm that automatic punctuation itself significantly increases understandability, even if several other factors interplay in subjective impression. The perceived improvement is even more pronounced in the DHH group. A statistical analysis is carried out to identify objectively measurable factors which are well reflected by subjective scores.

Index Terms: punctuation, subjective tests, RNN, low-latency, closed captioning

1. Introduction

Inserting punctuations into the transcripts provided by Automatic Speech Recognizers (ASR) has been a secondary task beside the efforts on lowering word error rates. Natural communication, however, implies that machines should be able to "write" what is spoken as a human could do, without telegraphic-style explicit dictation of the required punctuation marks. Although in some use-cases of ASR, punctuation may not be necessary at all – simple dialogue systems or voice control by commands do not require punctuation –, in use-cases such as transcription of meeting records, closed captioning, user friendly dictation etc., a proper and automatic insertion of punctuation marks can lead to significant improvement in the perceived "intelligence" and hence helpfulness of the system. The most challenging use-case is real-time large vocabulary ASR with punctuation (i.e. for example closed captioning), where a lightweight and low-latency punctuation module is required.

Automatic punctuation using sequence modelling principle and recurrent neural networks (RNN) yield good results recently [1, 2, 3, 4, 5] based on textual and/or acoustic (prosodic) features. However, part of these models rely on large context, including future context as well, which translates into high latency unsuitable for real-time exploitation. In [5] we proposed a lightweight low-latency punctuation model, and showed that only a modest performance decrease is associated with heavily limiting the future context of the punctuation model. We will use this framework in the present paper.

Any system is best evaluated by its end-users. As subjective testing may be time-consuming and expensive, objective measures are used for validation and testing, which can be also more carefully controlled by the objective requirements. A good objective fits well the subjective ratings and should preferably be easy and fast to evaluate and reproduce. In the domain of ASR, several studies addressed to predict the appropriateness of word error rate (WER) w.r.t. subjective ratings provided by ASR users [6, 7, 8, 9]. Obviously, not all ASR errors are equally disturbing or noticeable. A re-weighting of these errors based on syntactic information has been shown to increase correlation between WER and mean opinion scores (MOS).

Commonly used objective measures for automatic punctuation are borrowed from information retrieval: recall, precision or F-measure. The Slot Error Rate (SER) [10], inspired by the WER is also widely used to assess automatic punctuation. These measures seem quite technical and to the best of our knowledge, no attempt is documented on validating these measures by subjective tests. The interplay of word errors makes the picture more complex: it is reasonable to suppose, that word errors have higher impact, as punctuation provides primarily a structure [11] for the information contained in the words. Moreover, human error repair mechanisms [12, 13] may be able to mask the punctuation errors, especially as punctuation is supposed to be a less conscious process than correct spelling of words. Therefore, our primary research questions are whether (1) punctuation is a helpful cue in interpreting the meaning of a word chain which is not necessarily error free; and (2) whether the automatic punctuation, which itself is prone to errors, helps the reader at all? By the subjective evaluation it is essential to let the system to be scored by the primary target audience, that is deaf or hard-of-hearing (DHH) people.

This paper presents a subjective evaluation for the automatic punctuation module proposed in [5] used in an ASR which provides closed captioning for broadcast audio and video [14] in Hungarian language. Section 2 briefly presents our dataset and the RNN punctuation model. The subjective evaluation, including the test setups, is documented in Section 3. Finally, we provide a discussion and draw our conclusions.

2. Data and Method

2.1. The Hungarian Broadcast Dataset

The Hungarian dataset is provided by the Media Service Support and Asset Management Fund (MTVA) and covers broadcast video in various genres: weather forecasts, broadcast (BC) news and conversations, magazines, sport news and sport magazines. A subset with manual transcription and punctuation is used for training the RNN model [5]. The covered punctuation

marks include commas, periods, question marks and exclamation marks. Colons and semicolons are mapped to commas, all other punctuation marks are removed.

2.2. The Hungarian ASR

For the experiments we use the closed captioning system presented in [14]. WER on the entire punctuation test set was 24%, showing large variation depending on genre. More characteristics of the used subsets are summarized in [5]. We note that word error rates of the Hungarian ASR system are not directly comparable to the WERs of English ASR tasks due to the highly inflective nature of the language, a recognition error in a prefix or a suffix can make a whole word incorrect, hence WER tends to be higher for Hungarian than for English tasks, even if the subjective quality measures are close [15].

2.3. The RNN Punctuation Model

We use the model presented in [5]. The model gets a fixed-length word chain as input. Each word is projected into a semantic space with pre-trained word embeddings [16]. The following layer is a unidirectional LSTM layer to capture sequence as a context. The output is the predicted punctuation label obtained after a softmax activation of the last layer for the slot preceding the current word (the one before last in the sequence, as future context is limited to a single word). This simple structure allows for real-time operation with low-latency.

The vocabulary is limited to the 100K most frequent words in the training corpus, by mapping the remaining outliers to a shared *"Unknown"* symbol. During training on GPU, we use RMSProp optimizer, categorical cross-entropy loss and also let the imported embeddings to learn. An exhaustive objective evaluation of the model is presented in [5].

3. Subjective Evaluation Study

3.1. Test Procedure

Starting from the use-case of closed captioning, the goal is to transfer to the users in writing what is meant in speech. This is implemented by using an ASR to caption spoken content and convert it to text. Our interest is condensed around the following research questions: (Q1) to what extent is understandability of captions is influenced by the existence or lack of accurate (manual) punctuation marks; (Q2) is error-prone automatic punctuation still useful; and whether (Q3) can we separate the factors – ASR errors, punctuation errors, topic, etc. – which govern subjective opinion and are represented by a single score?

We compare six different captioning strategies as follows:

- MT-MP: Manual transcripts with manual punctuation
- AT-MP: ASR transcripts with manual punctuation
- MT-AP: Manual transcripts with automatic punctuation
- AT-AP: ASR transcripts with automatic punctuation
- MT-NP: Manual transcripts without any punctuation
- AT-NP: ASR transcripts without any punctuation

Combining the above 6 strategies and the 6 different genres (Section 2), 36 test sessions of 6-10 sentences (from the same context) are constructed in total, which are shuffled and presented to the users, who are instructed to read the text and rate it on a 5 grade scale as follows. 5: Excellent (Well understood, no errors perceived); 4: Good (Understood, some errors perceived); 3: Fair (After several reads finally understood); 2: Poor (Only partially understandable); 1: Bad (Not understandable). Aggregating these scores MOS is computed.

3.2. Test results

The subjective tests involved 181 participants (age: μ = 28.23 and σ = 9.20), 121 men and 60 women, leading to 460 ratings overall. Each subject rated at least 2 and at most 14 caption snippets. All subjects were native Hungarian speakers.

Mean Opinion Scores (MOS) for the different caption strategies (overall regarding genres) with pairwise Mann-Whitney U-tests (whether MOS are significantly different) are presented in Table 1[1].

Table 1: *MOS and pairwise Mann Whitney U-tests; * = significant by $p < 0.05$ with U-values in brackets.*

Caption Strategy	MT-MP	MT–AP	AT-MP	MT-NP	AT-AP	AT-NP	MOS
MT-MP	1						**4.27**
MT-AP	**0.002* (1874.5)**	1					**3.87**
AT-MP	**0* (1683.5)**	**0.007* (2318)**	1				**3.45**
MT-NP	**0* (1000.5)**	**0* (1435)**	**0.017* (2525)**	1			**3.13**
AT-AP	**0* (1286)**	**0* (1725.5)**	**0.013* (2771.5)**	0.436 (2896.5)	1		**3.02**
AT-NP	**0* (731.0)**	**0* (1063.5)**	**0* (2014.5)**	**0.033* (2247)**	0.376 (2828)	1	**2.84**

Fig. 1 shows the MOS values for each of the 6 genres with the 6 caption strategies.

3.2.1. Does punctuation help?

Obviously best MOS is seen with MT-MP in Table 1. For manual transcriptions (MT), both the manual and automatic punctuation were significantly more preferred over the unpunctuated strategy (MT-NP). In case of ASR transcriptions (AT), MOS of manually punctuated captions (AT-MP) is significantly higher not only to AT-AP and AT-NP but also to MT-NP, which means that even if the captions contain word errors, the presence of precise punctuation can counteract this and leads to better understandability. These findings suggest that people have a clear preference for punctuated texts.

Nevertheless, we could not report a significant difference between AT-AP and AT-NP, which basically constitute the two alternative use-cases available during automatic closed captioning with or without automatic punctuation. MOS is higher for AT-AP, but this difference is not significant. Analysing further these differences showed us significant WER dependency in the assessment of automatic punctuation as shown in Fig. 2. Indeed, higher WER often goes in hand with less formal speaking style. We can observe that when WER gets higher than a critical value, somewhere between 20% and 25% in our experiments, automatic punctuation has no benefits any more. If WER is below this threshold, that is for genres weather forecast, BC and sport news, AT-AP significantly outperforms AT-NP in MOS by $p < 0.05$. These are good news regarding the helpfulness of automatic punctuation.

3.2.2. How do word and punctuation errors interplay?

With spontaneous speaking styles (sport news and/or magazines) we indeed face the paradox that we intend to use punctuation, although spontaneous speech is by nature spoken and

[1] ANOVA is not applicable as we cannot assume the data normality.

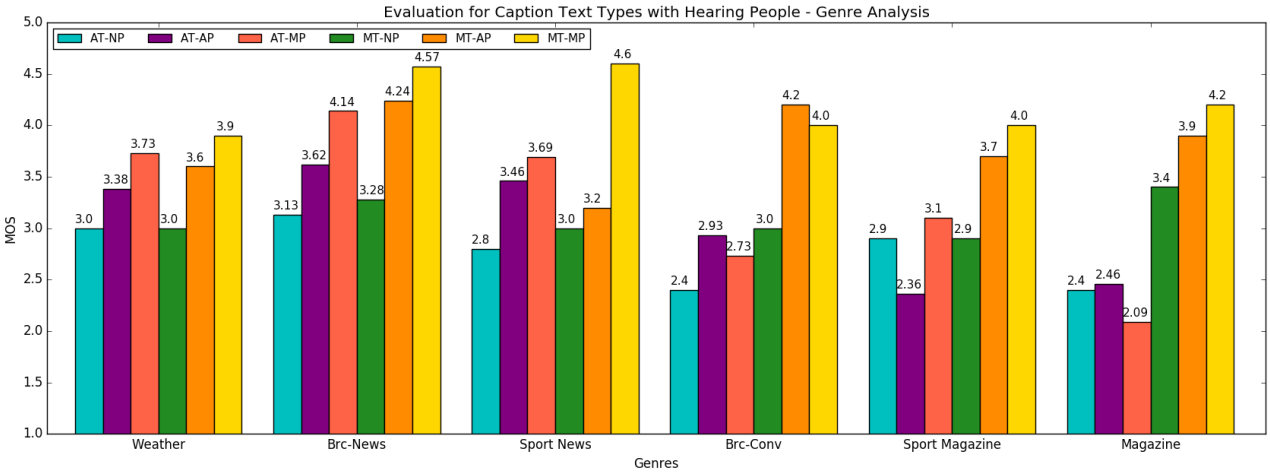

Figure 1: *MOS for the 6 caption strategies by genre*

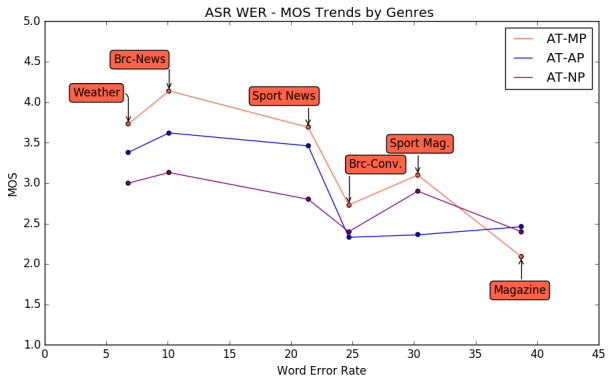

Figure 2: *WER - MOS trends for ASR-based caption strategies, labelled by genre*

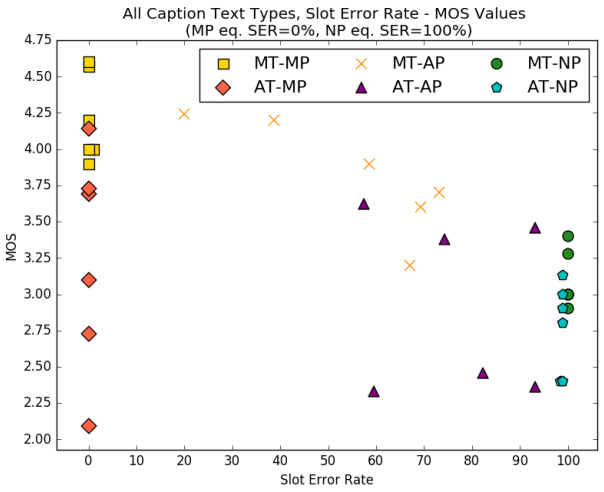

Figure 3: *SER-MOS plots for all caption strategies*

is not necessarily organized into sentences [17]. When using a reference to calculate SER, it is not any more coherent with the – erroneous – word sequence, and hence we should be very careful using SER to evaluate punctuation (c.f. [18]). Analysing MOS - SER plots shown in Fig. 3 provides us more insight into this aspect. In the MT-AP strategy (no word errors), Pearson correlation between SER and MOS is -0.39, which we can also observe in Fig. 3 suggesting a close to linear, albeit not strong relationship between SER and MOS when WER=0%. However, switching to the AT-AP strategy, which is the realistic use-case, this correlation could not be confirmed any more. Formal (higher MOS: weather forecast, BC and sport news) and spontaneous (lower MOS: BC and sport conversations, magazines) speaking styles are separated into the two observed clusters in Fig. 3. When both word and punctuation errors are present, SER was not informative at all regarding user rating.

3.2.3. Effect of punctuation errors on MOS

A user score depends on many factors. Supposing a part of these is determined by word and punctuation errors, a Generalized Additive Model (GAM) [19] can help in identifying the share that such factors X_i contribute to a user score Y. With GAM, the user score can be decomposed as follows:

$$g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_n(x_n), \qquad (1)$$

where $g(.)$ is the link function and $E[.]$ gives the expectation.

Defining $X_1 ... X_n$ such that they represent insertion, substitution and deletion errors for words and punctuation marks ($n = 6$) with smoothing splines estimates for $f_i(x_i)$, it turns out that punctuation insertion and substitution alone with the number of punctuation slots explain 32.3% of the variance observed in MOS. Deletion errors in punctuation were rare and hence we could not determine with sufficient certainty their contribution to MOS. Nevertheless, the higher impact of insertion errors in punctuation coincides with intuition, i.e. insertion errors were expected to be more disturbing as they provide a false structuring of the information, likely to counteract grammatical rules and constraints, whereas a deletion may be easier to recover by humans.

Table 2: *Joint (MT+AT) Closed Captioning results*

	AP vs. NP				AP vs. MP				MP vs. NP		
Genres	**AP**	**Same**	**NP**		**AP**	**Same**	**MP**		**MP**	**Same**	**NP**
Weather	**3**	**7**	**2**		**4**	**5**	**3**		**4**	**5**	**3**
Sport news	**9**	**1**	**2**		**3**	**8**	**1**		**8**	**4**	**0**
All (%)	**50.0**	**33.3**	**16.7**		**29.2**	**54.2**	**16.7**		**50.0**	**37.5**	**12.5**

3.3. Focus on DHH audience

In the previous sections we showed that hearing people tend to prefer punctuated captions, moreover they can profit from automatic punctuations as well if the quality of the ASR transcription exceeds a certain level in WER. We carried out subjective tests with DHH subjects to see how our former findings generalize to the primary audience of closed captioning.

We slightly changed the test setup to better reflect realistic usage conditions. 18 DHH students (aged 13-14 years) were asked to view short, 1-1,5 minute long coherent, muted and subtitled video recordings in a classroom experiment. We simulated static captioning, which means that the whole subtitle block is shown at once (one-shot appearance). Weather forecast and sport news samples were selected given the young age of our subjects.

Each video snippet was prepared with the same 6 captioning strategies, but instead of direct scoring, a comparative assessment was carried out: watching the same video pairwise in random order with different subtitling strategies, subjects performed comparisons on a prepared drawing, referring to a scale. They drew the arrow of the scale proportional to their subjective impression in favour of one of the videos regarding its understandability. Finally, this was quantized to three grades representing preference for either of the samples or a neutral opinion.

As we observed that ASR and manual transcriptions show the same tendencies, we present results comparing punctuation strategies only. Table 2 summarizes these results, showing a clear preference for punctuated captions (MP vs. NP, AP vs. NP), with an interesting, albeit not significant superiority of automatic punctuation over the manual one (AP vs. MP). The differences are more pronounced in videos related to sport news compared to weather forecasts. Pairwise exact tests [20] were performed, although in several cases the number of votes was not sufficient to conclude significant differences by p=0.05. Nevertheless, some significant differences could be confirmed:

1. AT-MP is significantly ($p = 0.048$) more preferred than AT-NP.
2. There is a significant ratio ($p = 0.012$) of votes (50% of the cases, 24 from 48), preferring the punctuated subtitles (MT-MP, MT-AP, AT-AP, AT-MP) versus the lack of punctuation (AT-NP+MT-NP).
3. For sport news, subjects were unable to make a difference between manual (MT-MP, AT-MP) and RNN punctuation (MT-AP+AT-AP); the number of votes reflecting neutral opinion on the difference is significantly higher than the two others (p=0.048).
4. For sport news experiments, there is a significant difference between the votes for captions with automatically restored punctuation marks (MT-AP+AT-AP) and captions without punctuation marks (MT-NP+AT-NP), favouring the punctuated one (p=0.012). On weather forecast captions the difference was not significant.

Examining the votes person by person, 61% of them (11/18) had a positive balance in favour of the enhanced captions (despite of some votes for unpunctuated subtitles), which means DHH people preferred punctuations in videos.

4. Conclusions

In this paper, we evaluated a low latency, RNN-based punctuation model, designed primarily for punctuation of closed captions. In [5] an exhaustive objective evaluation is run for this model, both for Hungarian and English. Here we focussed on subjective evaluation, i.e. whether punctuation adds a subjectively confirmed benefit to the captions, and what can we say about the relation between the used objective and subjective measures. We involved DHH subjects in order to represent the primary end-user audience of closed captioning.

The subjective evaluation process was designed such that it makes the assessment possible on word error-free transcripts (to evaluate clearly the share of punctuation in understanding a text) and ASR-produced transcripts (to test for realistic use-cases and to see whether punctuation keeps to be useful when word errors already degrade text quality). For punctuation, we compared three strategies: missing punctuation, error-free punctuation and machine produced punctuation.

Our results, obtained from a big sample survey, demonstrated clearly that users prefer punctuated text, even if punctuation is prone to some errors. MOS were significantly higher for RNN-punctuated texts, with the condition, that word errors occur up to a 20-25% WER (in Hungarian broadcast tasks). Indeed, it is easy to agree that once word errors trespass a critical amount, the punctuation task itself becomes hard to define, as the word chain to be punctuated is grammatically incorrect. A similar problem arises with spontaneous speech, where punctuation is not defined in the sense it is used in written language.

Beside significance tests on the obtained MOS for the 6 examined caption strategies, a GAM approach also confirmed that punctuation errors account for approx. 1/3 of the variance measured in the user scores.

Experiments with DHH subjects showed a more pronounced benefit in favour of punctuated captions, RNN-produced punctuation of ASR transcript was preferred over missing punctuation marks.

5. Acknowledgements

The authors would like to thank all volunteers taking part in the tests, especially the teachers and students of the Dr. Béla Török Kindergarten, Elementary School, Vocational School, Skills Development School, Unified Special Education Methodology Institute and Dormitory. The authors would like to thank the support of the Hungarian National Research, Development and Innovation Office (NKFIH) under contract ID *FK-124413*.

6. References

[1] E. Cho, J. Niehues, K. Kilgour, and A. Waibel, "Punctuation insertion for real-time spoken language translation," in *Proceedings of the Eleventh International Workshop on Spoken Language Translation*, 2015.

[2] O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," in *Proceedings of Interspeech*, 2016, pp. 3047–3051.

[3] V. Pahuja, A. Laha, S. Mirkin, V. Raykar, L. Kotlerman, and G. Lev, "Joint Learning of Correlated Sequence Labelling Tasks Using Bidirectional Recurrent Neural Networks," *arXiv preprint arXiv:1703.04650*, 2017.

[4] A. Moró and G. Szaszák, "A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery," in *Proceedings of Interspeech*, 2017.

[5] M. Á. Tündik, B. Tarján, and G. Szaszák, "Low Latency MaxEnt- and RNN-Based Word Sequence Models for Punctuation Restoration of Closed Caption Data," in *International Conference on Statistical Language and Speech Processing*. Springer, 2017, pp. 155–166.

[6] T. Monma, E. Sawamura, T. Fukushima, I. Maruyama, T. Ehara, and K. Shirai, "Automatic closed-caption production system on TV programs for hearing-impaired people," *Systems and Computers in Japan*, vol. 34, no. 13, pp. 71–82, 2003.

[7] R. Hong, M. Wang, M. Xu, S. Yan, and T.-S. Chua, "Dynamic captioning: video accessibility enhancement for hearing impairment," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 421–430.

[8] M. Federico and M. Furini, "Enhancing learning accessibility through fully automatic captioning," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. ACM, 2012, p. 40.

[9] S. Kawas, G. Karalis, T. Wen, and R. E. Ladner, "Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students," in *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2016, pp. 15–23.

[10] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*, 1999, pp. 249–252.

[11] E. Selkirk, "The syntax-phonology interface," in *International Encyclopaedia of the Social and Behavioural Sciences*. Oxford: Pergamon, 2001, pp. 15 407–15 412.

[12] A. Postma, "Detection of errors during speech production: A review of speech monitoring models," *Cognition*, vol. 77, no. 2, pp. 97–132, 2000.

[13] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[14] Á. Varga, B. Tarján, Z. Tobler, G. Szaszák, T. Fegyó, C. Bordás, and P. Mihajlik, "Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach," in *Proceedings of SPECOM*. Springer, 2015, pp. 105–112.

[15] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics -*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 487–494.

[16] M. Makrai, "Filtering Wiktionary triangles by linear mapping between distributed models," in *Proceedings of LREC*, 2016, pp. 2776–2770.

[17] F. Goldman-Eisler, "Pauses, clauses, sentences," *Language and speech*, vol. 15, no. 2, pp. 103–113, 1972.

[18] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[19] T. Hastie and R. Tibshirani, *Generalized additive models*. Wiley Online Library, 1990.

[20] C. R. Mehta and N. R. Patel, "IBM SPSS exact tests," *SPSS Inc., Cambridge, MA*, 2010.