SENTENCE BOUNDARIES IN TEXT AND PAUSES IN SPEECH: CORRELATION OR CONFRONTATION?

Anton Stepikhov[a], Anastassia Loukina[b]

[a] St. Petersburg State University, Russia; [b] Educational Testing Service, USA
a.stepikhov@spbu.ru; aloukina@ets.org

ABSTRACT

The paper explores the interaction between sentence boundaries marked by annotators in transcriptions of Russian spontaneous speech and actual prosodic boundaries in the signal. The aim of the research is to investigate whether annotators' prosodic competence allows them to correctly detect sentence boundaries in speech based on textual information only.

We found that inter-annotator agreement for each sentence boundary identified in transcription was affected by both presence or absence of pause and pause duration. Mixed linear model showed that presence or absence of pause explain 13% of variance in boundary detection. Pause duration explained only 4% of variance in inter-annotator agreement with moderate correlation of $r = 0.21$.

We argue that relatively small size of effect in this case may be due to the interaction of different pausing strategies typical for reading and spontaneous speech, ambiguity of sentence boundaries and individual differences in speech perception.

Keywords: boundary detection, pausing, annotation, spontaneous speech, Russian.

1. INTRODUCTION

The problem of the segmentation of unscripted speech into sentences for decades has been one of the key issues for both linguistics and computer science [27, 10, 20, 6]. Speech, unlike a written text, does not specifically mark sentence boundaries and therefore cannot be segmented into sentences unambiguously [19, 16, 21, 17, 21].

Nonetheless, as discussed in [23] linguists need a unit for further analysis [cf. 7], and therefore they have to face the problem of oral text segmentation. "Does a sentence exist in speech?" – this is the question they try to answer [13]. Though sometimes preference might be given to non-sentence units (e.g. *elementary discourse units* in [14]), the notion of sentence is still viable for linguistic analysis. Acquiring information on sentence boundaries is also crucial for natural language processing and automatic speech recognition as it improves language processing techniques and enhances human readability of recognised speech [20, 10].

To identify sentence boundaries in unscripted speech, expert manual annotation is usually used. It is considered to be the gold standard for sentence boundary detection. Manual annotation can be based on textual and prosodic information, but the interaction between the two is not yet fully understood. For example, for Russian speech [27] showed that the influence of the semantic factors on segmentation outweighs that of the prosodic factors. The analysis of sentence boundary detection in a Russian ASR system revealed that in Russian spontaneous speech it is difficult to detect boundaries based on prosodic clues alone [6].

While most previous studies provided annotators with both prosodic and textual information, we earlier explored whether sentence boundaries can be reliably detected based on text transcription alone [23-24]. We suggested that when experts have no access to the actual recordings they can focus on semantic and syntactic features of the text and showed that the annotators could reach relatively high agreement in boundary detection based on textual information only – without drawing upon prosody.

In [24] we argued that such text-based sentence boundaries reflected segmentation in inner speech of the annotator as she or he is reading the transcription. Thus, the lack of information about a speaker's intonation is to some extent compensated by the reader's prosodic competence, allowing him or her to feel the rhythm and melody of sentences without hearing the actual speech [8].

In this paper we question this assumption and explore to which extent boundaries guessed from textual transcription of speech correspond to actual prosodic boundaries in the signal. Does prosodic competence really work in this case and allow the reader to reconstruct the intonation of a speaker they never heard from the written transcription of this speaker's speech?

To answer this question, we conducted a pilot research based on the same corpus of Russian spontaneous monologue as used in [23-24] and compared two types of annotation: expert manual annotation of unscripted speech based on textual

information and prosodic annotation. We used pauses as markers of prosodic boundaries since pause is known to be one of the acoustic boundary marks both in Russian and many other languages, e.g. Danish, Swedish, English, European Portuguese, French, Finnish, Thai [cf. 26, 11, 12, 4]. Based on these data, we performed statistical analysis to find out whether there is a correlation between expert's estimation of a sentence end and a real pause in these positions.

2. DATA AND METHOD DESCRIPTION

2.1. Corpus

The study is based on the corpus of spontaneous Russian monologues described in [23]. This corpus contains manual transcriptions of different types of monologues recorded by 32 native speakers of Russian. Each speaker was presented with several tasks: (1) to read a story with a plot and subsequently retell it from memory ('story'), (2) to read a descriptive narrative without a plot and retell it from memory ('description'), (3) to describe a series of pictures in a cartoon ('picture story'), (4) to describe a landscape painting ('picture description'), and finally (5) to comment on one of the suggested topics ('free comment').

Thus, the corpus consists of 160 texts (~55k words), with overall duration about 9 hours. The corpus data is balanced with respect to speakers' social characteristics (e.g. gender, age, use of speech in everyday life) and text types.

2.2. Texts for the analysis

For this pilot study, we used two types of monologues from the corpus – picture story and free comment, since these types of text are opposed in terms of inter-annotator agreement. As we showed in [23], the agreement is highest for picture story and lowest for free comment.

We investigated 12 texts produced by 6 male speakers (2 texts by each). The total duration of the analysed monologues is 45 minutes (11.7 min for picture story and 33.4 min for free comment) or 5,952 words (1,305 words for picture story and 4,647 – for free comment). Summary statistics is shown in Table 1.

Table 1: Summary statistics of analysed texts by text type.

Type of text	Duration		Words	
	Mean (sec)	SD	Mean (count)	SD
Picture story	116.8	72.9	217.5	120.8
Free comment	334.3	287.4	774.5	708.4

2.3. Expert manual annotation

The corpus of Russian spontaneous monologues also includes manual annotations of sentence boundaries. These were collected using orthographic transcriptions of recorded speech (see [23] for further detail). The transcription did not contain any punctuation. To make text reading and perception easier, graphic symbols of hesitation (like *eh*, *uhm*) and other comments (e.g. [sigh], [laughter]) were also excluded.

These transcripts were then manually segmented into sentences by a group of experts consisting of 20 native speakers of Russian with a background in linguistics who were asked to mark sentence boundaries using conventional full stops or any other symbol of their choice (e.g. a slash). The annotation was performed based on textual information only. The experts were presumed to have a native intuition of what a sentence is and, thus, it was left undefined. There were no time-constraints.

2.4. Prosodic annotation

For prosodic annotation, we identified all positions in the transcriptions where at least one annotator marked a sentence boundary. We then used Wave Assistant software [28] to identify whether the actual recording contained a pause in those positions and if so what was the duration of the pause. Hesitations were considered to be part of the pause. This annotation was done manually by an expert in Russian phonetics.

3. DATA ANALYSIS

Following [24] for each position in the text we computed the number of experts who had marked the boundary at this position. This number is interpreted as a "boundary confidence score" (BCS) which ranges from 0 (no boundary marked by any of the experts) to 20 (boundary marked by all experts = 100% confidence).

The total amount of positions with BCS > 0 in the analysed texts is 1333: 227 positions in picture stories and 1106 positions in free comments.

3.1. Difference between types of text

Table 2 shows average BCS and average duration of corresponding pauses in each type of text.

Table 2: Summary statistics of analysed text types with regard to BCS and pause length.

	Picture story		**Free comment**	
	BCS	Pause duration (ms)	BCS	Pause duration (ms)
Median	8	358	4	0
Mean	8.65	623	5.88	367
SD	6.43	769	5	633
N	227		1106	

Since our data included multiple measurements from the same speaker which cannot be considered independent, we used hierarchical (mixed-effects) linear models [2] to evaluate the connection between BCS and text type, presence or absence of pause and the duration of the pause. All models were fitted using [3] with *p*-values estimated using [16].

Mixed linear models with BCS or pause length as dependent variables, type of text as fixed factor and speaker as random factor showed that both pause duration and BCS were significantly higher in picture story than in free comment ($p < 0.0001$). After controlling for between-speaker variation, the difference in average BCS between picture story and free comment was 2.15 (9.03 vs. 6.88), while the difference in average pause duration was 250 ms. In other words, monologues describing sequence of pictures contained longer pauses in positions marked as sentence boundaries in transcriptions and had higher inter-annotator agreement about the position of sentence boundaries than free comment monologues. Note that in free comment more than half of all BCS corresponded to no pause at all (624 or 56%). In picture story there were 87 such positions (38%). At the same time, the type of text accounted for only a small share of variance of both pause duration and BCS: analysis of explained variance showed that type of text explained about 4% of variance in BCS and only about 2% of variance in pause length (we followed the same procedure as in [19]).

3.2. BCS and presence or absence of pause

We next explored whether there was significant difference in BCS depending on presence or absence of a pause. We used the same models as in 3.1 but added a binary variable indicating presence or absence of pause as another fixed factor. Analysis of likelihood showed that this has led to significant improvement in model fit ($p < 0.0001$). While the main effect of text remained significant after the addition of pause variable ($p < 0.0001$), we found no further interaction between text type and presence of pause ($p = 0.618$). After controlling for between-speaker variability, the difference between average BCS in positions with or without pause was 10.28 vs. 6.54 for picture story and 8.82 vs. 5.08 for free comment. Furthermore, presence or absence of pause explained further 13% of variation in BCS.

Table 3 shows the final estimates of BCS for different text types and positions after controlling for within-speaker variation.

Table 3: Average BCS for different text types and positions as estimated by mixed linear model with speaker as random factor.

	Picture story	Free comment
Pause	10.28	8.82
No pause	6.54	5.08

3.3. BCS and pause duration

Finally, we explored the connection between the strength of the boundary and the duration of the pause. For this analysis we only used 627 boundaries where the pause duration was not equal to zero. Table 4 shows summary statistics for this subset of data.

Table 4: Summary statistics of analysed text types with regard to BCS and pause length in positions with pause.

	Picture story		**Free comment**	
	BCS	Pause duration (ms)	BCS	Pause duration (ms)
Median	10	794	7	695
Mean	10.15	972	8	841
SD	6.3	762	5.2	721
N	145		182	

As already indicated in Table 3, the difference between text types observed on the whole data set was still significant for this subset ($p = 0.008$), although the gap became smaller: after controlling for within-speaker variability average BCS was 10.3 in picture story and 8.9 in free comment. The difference between texts explained 2% of variation in BCS in this subset of data.

Since both pause duration and BCS varied substantially between speakers, we standardised these values within each recording by using z-scores.

Our analysis showed that after standardisation the data was no longer clustered by speaker and therefore it was no longer necessary to use mixed model to account for between speaker variability. Since the data was standardised within each recording, there also was no difference between different types of text.

We found that there was significant but very weak correlation between standardised pause duration and BCS: Pearson's $r = 0.21$, $p < 0.0001$. Linear model with BCS as dependent variable and pause duration as independent variable showed that pause duration accounted for about only 4% of variability in BCS ($F_{(1, 625)} = 29.36$, *adj*. $r^2 = 0.04$, $p < 0.0001$).

4. DISCUSSION AND CONCLUSIONS

In this paper we explore whether speakers' prosodic competence allows them to reconstruct pauses in spontaneous speech based on transcription only without hearing the speech itself. To achieve our goal, we compared (1) whether sentence boundaries marked by expert annotators in textual transcriptions of Russian spontaneous speech corresponded to pauses in speech in actual recording; (2) whether boundary confidence score (BCS), i.e. the number of annotators who marked sentence boundary in a given position, is correlated with the duration of the pause.

We found that both presence or absence of pause and pause duration have statistically significant effect on BCS; however, the size of that effect remained relatively small. Mixed linear model showed that presence or absence of pause explain 13% of variance in BCS. The effect of pause duration was much weaker: for positions where pause is present, pause duration explained only 4% of variance in BCS with moderate correlation of $r = 0.21$.

One likely explanation is that in Russian, as in other languages, the pause is not the only cue to prosodic segmentation. Although pauses are often used as a convenient way to establish boundaries between prosodic units (cf. [15, 1]), prosodic boundaries can also be indicated by other acoustic cues such as pitch movement, intensity or duration of preceding segments. Sometimes these may not be accompanied by a pause [26, 12]. Since our analysis was limited to pause length, it does not take into account prosodic boundaries indicated by other acoustic cues. However, strong prosodic boundaries are usually indicated by the combination of all prosodic cues while our results showed that even boundaries with very high inter-annotator agreement were not necessarily accompanied by a pause.

Another reason for this discrepancy between boundaries identified in written text and pauses in speech is different functional load of pauses in various types of speech. As [5] pointed out, the number of "zero" pauses at the boundaries of the intonation units in Russian spontaneous speech is greater compared to reading. [9] showed that pausing in reading of English spontaneous speech differs from pausing in the original speech. We suggest that when the annotators were identifying sentence boundaries in transcriptions, they might have been using segmentation strategies common for reading rather than spontaneous speech. This in turn may have lead to the discrepancy between 'imagined' pauses in inner speech and real pauses.

Finally, as discussed in the introduction, prosodic boundaries in spontaneous speech are regularly ambiguous and therefore often depend on annotator individuality (cf. [25]). The process of annotation is inextricably linked with speech perception, understanding and interpretation. Lastly, grammatical structure of the text may be ambiguous and allow different interpretations [24].

At the same time, our results showed that there was significant difference between boundary confidence scores depending on presence or absence of pause in the recording. In other words, more annotators marked sentence boundaries in places where the speaker made a pause. Therefore the annotators are able to use their prosodic competence to correctly identify at least some of the prosodic boundaries. Thus, we argue that there are both correlation and confrontation between sentence boundaries in annotated texts and pauses in real speech. Correlation is explained by annotators' prosodic and – wider – communicative competence. In its turn, confrontation may be determined, first, by interaction of different pausing strategies typical for reading and spontaneous speech, second, by ambiguity of sentence boundaries and, finally, by individual speech perception.

Our study was based on a relatively small corpus and therefore the results should be taken with caution. In future we plan to expand this research to a large dataset as well as explore how the annotation correlates with other prosodic factors.

ACKNOWLEDGEMENTS

This study was supported by the Russian Foundation for Humanities, project No. 15-04-00165. We thank Lei Chen, Keelan Evanini, and Su-Youn Yoon for their comments and suggestions.

5. REFERENCES

[1] Aylett, M., Turk, A. 2004. The Smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47 (1), 31–56.

[2] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59 (4), 390–412.

[3] Bates, D., Maechler, M., Bolker, B., Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-5. http://CRAN.R-project.org/package=lme4.

[4] Bolinger, D. 1998. Intonation in American English. In: Hirst, D., Di Cristo, A. (eds), *Intonation Systems. A Survey of Twenty Languages*, Cambridge: Cambridge University Press, 45–55.

[5] Bondarko, L. V., Volskaya, N. B., Vasilyeva, L. A., Tananayko, S. O. 2003. On phonetic properties of russian read and spontaneous speech. *Proc. 15th ICPhS* Barselona, 2973–2976.

[6] Chistikov, P., Khomitsevich, O. 2011. Online automatic sentence boundary detection in a Russian ASR system. *SPECOM 2011. The 14th International Conference "Speech and Computer"*. Kazan, 112–117.

[7] Foster, P., Tonkyn, A.,Wigglesworth, G. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21(3), 354–375.

[8] Gasparov, B. M. 1996. *Yazyk, pamyat', obraz. Lingvistika yazykovogo sushchestvovaniya*. Moscow: Novoe literaturnoe obozrenie. (in Russian)

[9] Goldman-Eisler, F. 1972. Pauses, clauses, sentences. *Language and Speech* 15 (2), 103-113.

[10] Gotoh, Y., Renals, S. 2000. Sentence boundary detection in broadcast speech transcripts. *Automatic Speech Recognition: Challenges for the new Millenium, ISCA Tutorial and Research Workshop (ITRW)*, Paris, 228–235.

[11] Grønnum, N. 1998. Intonation in Danish. In: Hirst, D., Di Cristo, A. (eds), *Intonation systems. A survey of twenty languages*, Cambridge: Cambridge University Press, 131–151.

[12] Hirst, D., Di Cristo, A. 1998. A survey of intonation systems. In: Hirst, D., Di Cristo, A. (eds), *Intonation systems. A survey of twenty languages*, Cambridge: Cambridge University Press, 1–44.

[13] Kibrik, A. A. 2008. Est' li predlozhenie v russkoj rechi? In: Arkhipov, A. V., Zakharov, L. M., Kibrik, A. A., et al. (eds.), *Phonetics and non-phonetics: For the 70th birthday of Sandro V. Kodzasov.* Moscow: Jazyki slavjanskih kul'tur, 104–115. (in Russian)

[14] Kibrik, A. A., Podlesskaya, V. I. (eds.) 2009. *Night dream stories: A Corpus Study of Spoken Russian Discourse*. Moscow: Jazyki slavjanskih kul'tur. (in Russian)

[15] Kochanski, G., Shih, C., Jing, H. 2003. Quantitative measurement of prosodic strength in Mandarin. *Speech Communication* 41(4), 625–645.

[16] Kuznetsova, A., Brockhott, P. B., Christensen, R. H. B. 2014. lmerTest: Tests in Linear mixed-effects models. R package version 2.0-20. http://CRAN.R-project.org/package=lmerTest.

[17] Liu, Y. 2004. *Structural event detection for rich transcription of speech*. PhD thesis. Purdue University.

[18] Liu, Y., Chawla, V. N., Harper, M. P., Shriberg, E., Stolcke, A. 2006. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech and Language* 20 (4), 468–494.

[19] Loukina, A., Rosner, B., Kochanski, G., Keane, E. 2013. What determines duration-based rhythm measures: text or speaker? *Laboratory Phonology*, 4 (2), 339–382.

[20] Nasukawa, T., Punjani, D., Roy, S., Subramaniam, L. V., Takeuchi, H. 2007. Adding sentence boundaries to conversational speech transcriptions using noisily labelled examples. *AND* 2007, 71–78.

[21] Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tur, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., Woofers, C. 2008. Speech segmentation and spoken document processing. *Signal Processing Magazine, IEEE*, 25 (3), 59–69.

[22] Skrebnev, Yu. M. 1985. *Vvedenie v kollokvialistiku.* Saratov: Izdatel'stvo Saratovskogo universiteta. (in Russian)

[23] Stepikhov, A. 2013. Analysis of expert manual annotation of the Russian spontaneous monologue: Evidence from sentence boundary detection. *Proc. SPECOM 2013. LNCS (LNAI)* 8113, 33–40.

[24] Stepikhov, A. 2013. Resolving ambiguities in sentence boundary detection in Russian spontaneous speech. *Proc. TSD 2013. LNCS (LNAI)* 8082, 426–433.

[25] Stepikhov, A., Loukina A. 2014. Annotation and personality: Individual differences in sentence boundary detection. *Proc. SPECOM 2014. LNCS (LNAI)* 8773, 105–112.

[26] Svetozarova, N. 1998. Intonation in Russian. In: Hirst, D., Di Cristo, A. (eds), *Intonation systems. A survey of twenty languages*, Cambridge: Cambridge University Press, 264–277.

[27] Vannikov, Yu., Abdalyan, I. 1973. Eksperimental'noe issledovanie chleneniya razgovornoj rechi na diskretnye intonacionno-smyslovye edinicy (frazy). In: Sirotinina, O. B., Barannikova, L. I., Serdobintsev, L. Ja. (eds.), *Russkaya razgovornaya rech*. Saratov: Izdatel'stvo Saratovskogo universiteta, 40–46. (in Russian)

[28] *Wave Assistant 2.50*. Phonetic annotation software developed and distributed by Speech Technology Centre, Russia (www.speechpro.ru).