Not all polar questions are the same: ASR, Humans, and Russian

Maria Onoeva
Charles University and Humboldt-Universität zu Berlin
onoevam@ff.cuni.cz

Abstract

Word Error Rate (WER) remains the standard metric in automatic speech recognition (ASR) evaluation, yet it does not capture higher-level linguistic distinctions such as prosody. This article examines how three state-of-the-art open-source ASR models (Whisper, Meta's MMS, and GigaAM) handle the distinction between Russian polar questions and assertions. Russian is particularly suitable for this investigation because polar questions can be marked either morphologically (*li, razve*) or purely intonationally, without changes in word order. Using audio stimuli from a controlled psycholinguistic experiment, I compare human classification performance in two experimental studies with ASR transcriptions, taking sentence-final punctuation as a proxy for prosodic interpretation. While human participants show near-ceiling accuracy, the ASR models perform inconsistently, especially on intonationally marked questions. Additional contextual cues improve performance in some cases but also reveal instability across conditions. The results demonstrate that evaluating punctuation provides insights beyond WER and allows a more fine-grained view of how current ASR systems encode prosodic and grammatical information.

Keywords: ASR, Russian, polar questions, intonation

1. Introduction

Moving beyond Word Error Rate (WER) as the sole evaluation metric for automatic speech recognition (ASR) systems is essential for a more comprehensive assessment of their performance (Aksënova et al., 2021; Gandhi et al., 2022). While WER captures lexical accuracy, it does not reflect higher-level linguistic properties. In many languages, crucial grammatical distinctions are expressed through intonation rather than through segmental morphology or word order (e.g., focus or question marking; Ladd 2008, p. 5). However, despite operating on acoustic input, most ASR systems are optimized for lexical transcription rather than for the explicit modeling of suprasegmental structure (Renals and King, 2010, p. 814, 829).

In this paper, I investigate the extent to which three off-the-shelf ASR models encode prosodic distinctions in a low-resource experimental setting. Following earlier work that considers punctuation as an additional evaluation dimension (Meister et al., 2023; Gris et al., 2023), I use sentence-final punctuation as a proxy for prosodic interpretation. While punctuation does not map to speech in a one-to-one manner, it provides a measurable way to probe whether ASR systems differentiate between sentence types that are structurally identical but prosodically distinct.

Using audio stimuli from a psycholinguistic eye-tracking study designed to examine the processing of Russian polar (yes/no) questions and assertions (Razguliaeva et al., to appear), I compare human classification performance with the transcriptions produced by three state-of-the-art open-source models —OpenAI's Whisper (Radford et al., 2022), Meta's Massively Multilingual Speech (MMS) (Pratap et al., 2023), and GigaAM (Kutsakov et al., 2025). Russian offers a particularly suitable testing ground, as polar questions may be marked either morphologically or purely intonationally, without changes in word order. This allows us to directly assess whether current ASR systems capture distinctions that rely exclusively on prosody.

This paper is organized as follows. Section 2 introduces the relevant properties of Russian polar questions and assertions. Section 3 presents the ASR models and experimental results. Section 4 discusses the findings, and Section 5 concludes.

2. Russian PQs and assertions

Classified as a "Question Particle" language in Dryer (2013), Russian, in fact, exhibits two polar question (henceforth, PQ) strategies, both of which allow negation (Restan 1972; Zanon 2024; Korotkova 2023; Šimík to appear, a.m.o.): (i) overt particle marking as in (1) or (2), and (ii) intonational marking as in (3). The particle *li* attaches to a fronted verb in canonical matrix PQs as in (1) (King, 1994). PQs with another question particle *razve*, as in (2), are dubbed biased and approximately translated to English PQs with 'really' (Geist and Repp 2023, cf. Korotkova submitted).

(1) (Ne) Zažëg li Miša večerom svečku?
NEG lit LI Miša evening candle
'Did Miša (not) light a candle in the evening?'

(2) Razve Kira (ne) xodila segonja v školu?
RAZVE Kira NEG went today in school
'Did Kira really (not) go to school today?'

Proceedings of SPEAKABLE @ LREC 2026, pages 31–38
11 May 2026.

The latter intonation strategy in (3) preserves declarative SVO order with questionhood indicated exclusively by prosody, placing a special nuclear pitch accent L+H* on the inflected verb (Meyer 2004; Esipova 2025).[1] Compare an IntonPQ in (3) that carries the prominence on the verb *zažëg* 'he lit' (visual representation in Figure 1) with an assertion in (4) (also in Figure 2).[2] Meyer and Mleinek (2006, p. 1616) point out that this type of PQs might sound impolite or rough to English or German ears; however, on par with LiPQs, IntonPQs are attested in out-of-the-blue contexts, i.e., they are deemed canonical in Russian. But contrary to LiPQs, IntonPQs are more frequent in spoken speech and corpora (Restan 1972; Bryzgunova 1975; Esipova 2025; King 1994; Onoeva and Staňková 2025).

(3) Miša (ne) zažë$_{L+H*}$g večerom svečku?
Miša NEG lit evening candle
'Did Miša (not) light a candle in the evening?'

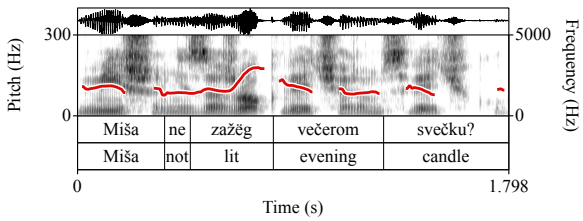

Figure 1: IntonPQ in (3)

(4) Miša (ne) zažëg večerom lampočku.
Miša NEG lit evening lightbulb
'Miša did (not) light a lightbulb in the evening.'

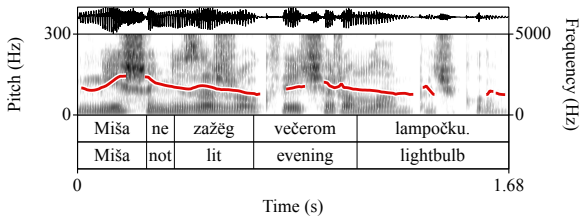

Figure 2: ASSERTION in (4)

As expected, native Russian speakers presented with just auditory cues (≈ 1 to 2 seconds recorded by a male speaker) detect the differences between the two sentence types at a very high level. During the eye-tracking visual world paradigm experiment reported in Razguliaeva et al. (to appear), we were able to observe the differences in processing between assertions like (4) (16 positive and 16 negative sentences) and PQs (32 positive and 32 negative for each LiPQs in (1) and IntonPQs in (3)): while in assertions, participants (N = 52) focused their attention on the picture corresponding to the expressed polarity, i.e., to a picture with the unlit lightbulb for (4) with negation and to the lit one with a positive assertion. Fixations in questions became concentrated on the positive picture (i.e., the lit candle for (1) and (3)), largely independently of the polarity expressed in the stimulus. The results were further replicated in a forced-choice task conducted with the same audio stimuli. The task was to listen to an audio recording and select one of the two presented options: for the sentences as in (1), (2), and (3), participants (N = 54) were expected to press "question," while for (4) they were to press "assertion". Participants showed near-ceiling performance in utterance classification; the descriptive results are in Figure 3 (means $= 0.97$-1, SD $= 0.00$-0.16, with little variance for statistical models to detect group differences).[3]

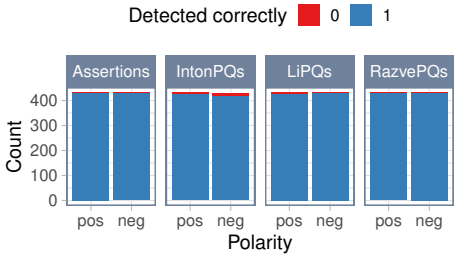

Figure 3: Plain audios results: Human accuracy.

The same task, when performed on the identical set of audio files by the three ASR models, proved to be considerably more problematic, as shown below.

3. ASR and Russian PQs

The evaluated ASR systems differ in architecture and decoding strategy. Whisper (Radford et al., 2022) is an autoregressive encoder–decoder model, `large_v3` was used for the experiments; transcriptions were generated using greedy decoding (temperature $= 0.0$). MMS by Meta (Pratap et al., 2023) is a CTC-based end-to-end model decoded via standard greedy CTC decoding. Since MMS does not generate punctuation, I restored punctuation in a post-processing step using the

[1] Alternatively, this nuclear pitch accent can be placed on the linearly last stressed syllable, resulting in an explanation-seeking question with a higher Question Under Discussion (Esipova and Romero, 2023; Esipova, 2025). It is not examined here and is left for future research.

[2] The graphics are compiled in Praat (Boersma and Weenink, 2009).

[3] All statistical analyzes and visualizations were performed in R (R Core Team, 2021).

Table 1: Plain audios: Humans and ASR models accuracy (%)

	ASSERTIONS		INTONPQS		LIPQS		RAZVEPQS	
	neg	pos	neg	pos	neg	pos	neg	pos
Humans (N=54)	100	99.8	97.2	99.1	99.8	98.6	99.8	100
Meta's MMS	100	100	0	0	18.8	18.8	68.8	68.8
GigaAM (CTC)	100	100	96.9	56.2	100	93.8	100	100
GigaAM (RNNT)	100	100	100	34.4	100	90.6	100	100
Whisper	100	100	62.5	6.2	96.8	84.4	100	100

external Silero Text Enhancement model (Silero Team, 2026). GigaAM (Kutsakov et al., 2025), designed specifically for Russian, was evaluated in two variants: (i) a CTC model decoded with greedy CTC decoding and (ii) an RNNT model using its standard streaming decoding. Except for Silero, no external language models were applied, ensuring comparability under low-resource conditions.

For Russian speech, the WER reported for the base Whisper `large_v3` model is 5.8 % on Common Voice and 5 % on FLEURS (OpenAI, 2022). The end-to-end variant of GigaAM-v3, which was used for both CTC and RNNT, won 70 % of pairwise side-by-side comparisons against Whisper (Salute Developers, 2023). MMS 1B achieves approximately 15 % WER on Russian in the multilingual FLEURS evaluation (Pratap et al., 2023). However, for the present study, punctuation, namely, a period for assertions and a question mark for questions, was used as a dependent variable in the ASR experiments.

3.1. Plain audios – results

Accuracy for the ASR models was defined as the proportion of utterances whose sentence-final punctuation matched the intended sentence type (question mark for questions, period for assertions). All other outputs were counted as incorrect. Similar to humans, detecting plain assertions and RAZVEPQS posed no difficulty for the ASR models, with near-ceiling accuracy across polarity conditions, see Table 1 and Figure 4 for the results. The only exception was MMS, which reached 68.8 % for RAZVEPQ. A larger variation emerged for LIPQS, while MMS showed low accuracy, GigaAM (CTC and RNNT) performed strongly on these questions, and Whisper remained slightly below ceiling.

The most striking differences appeared for INTONPQS: unlike top-performing humans, the models often fail to differentiate between the sentence types. MMS did so completely with 0 % accuracy, meaning it placed the period instead of the question in all INTONPQS. GigaAM models showed different results on the same set of audios: while RNNT reached 100 % accuracy in the negative condition, it dropped sharply in the positive one (34.4 %), resulting in a stronger polarity asymmetry compared to

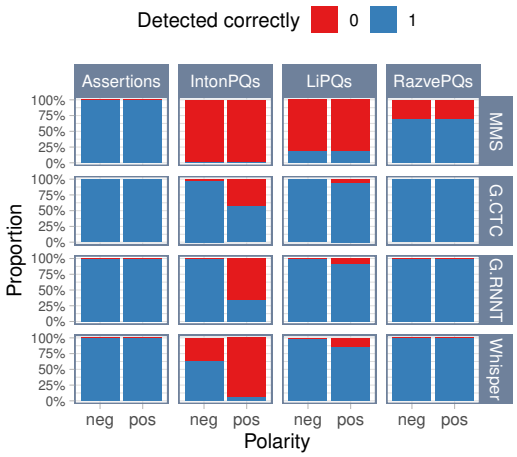

Figure 4: Plain audios results: ASR accuracy.

CTC (96.9 % vs. 56.2 %). Whisper also performed inconsistently, detecting only 6.2 % of positive INTONPQS and 62.5 % of negative. Overall, models handled morphologically marked questions better than purely intonational ones. To further investigate whether transcription accuracy can be improved, I conducted additional tests with the same audio stimuli enriched by some context.

3.2. Enriched audios – results

The original audio files recorded by a native male Russian speaker were enriched with nine additional contexts spoken by a female native speaker. The contexts were either related to questioning or unrelated, and were added before or after the plain utterances. Question-related contexts included polar replies such as *Ja dumaju, čto da/net* 'I think that yes/no' appended after a sentence, and preposed cues such as *On sprosil* 'He asked' and *Sledujuščee predloženie — èto vopros* 'The next sentence is a question'. Unrelated contexts included the prepended *On skazal* 'He said', the word *čerepaxa* 'turtle', and excerpts from Pushkin's *Ruslan and Ljudmila* (R&L), added either before or after the original audio. Figure 5 summarizes the results of all ASR experiments across four models.

Each heatmap shows accuracy (%), with darker shades indicating higher performance. Rows rep-

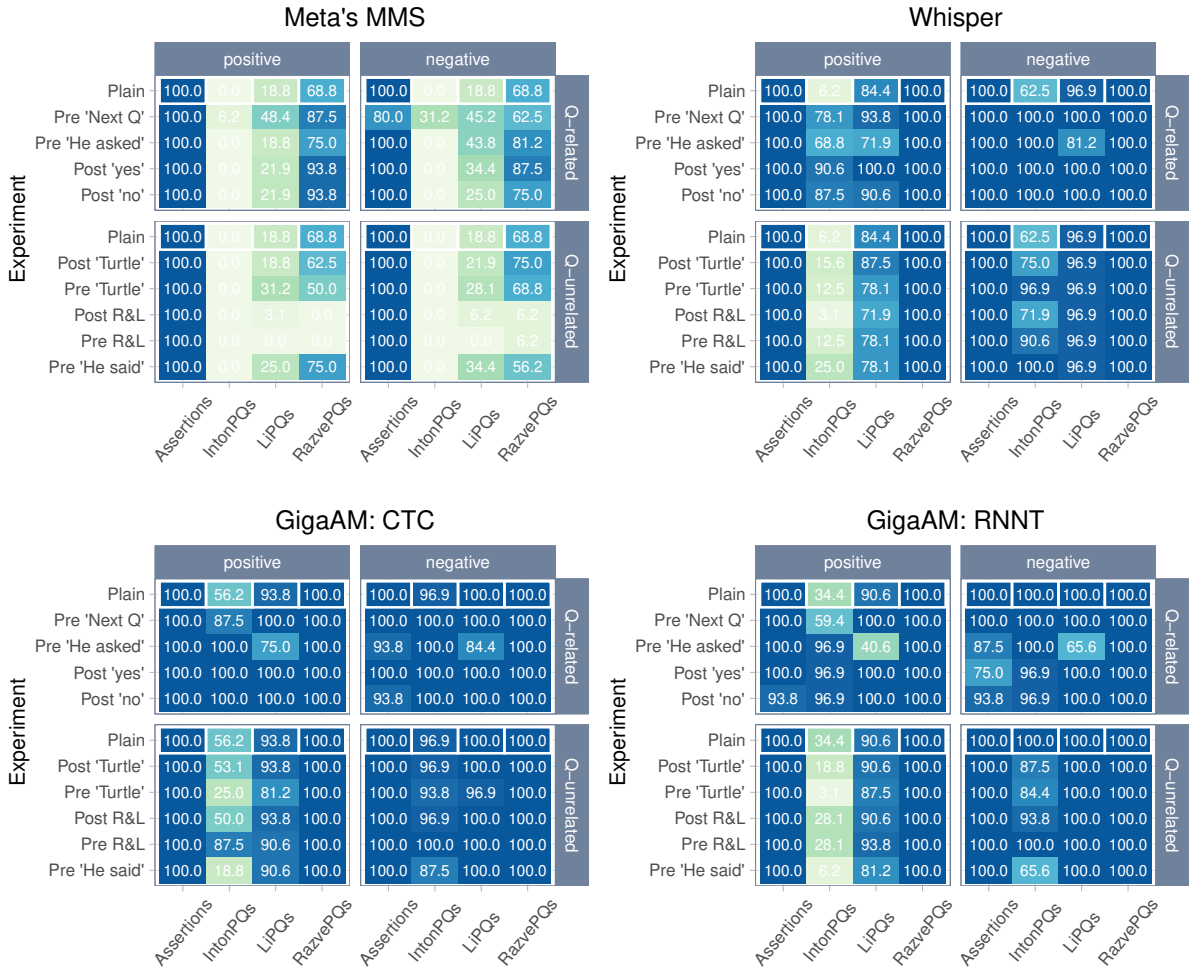

Figure 5: All ASR experiments: darker color = higher accuracy (%). The first row condition PLAIN had no context and added to each facet as a baseline comparison.

resent context manipulations, while columns distinguish between sentence types with and without negation. The first row in each facet corresponds to plain audio (i.e., from Table 1) and serves as a no-context baseline. Compared to the plain audios, the additional utterances had virtually no effect on assertions and RAZVEPQs (the first and last columns, respectively), which remained at or near ceiling across models and contextual manipulations. Once again, except for Meta's MMS in the preceding and added after lines from R&L, which, however, already had low accuracy in the baseline.

The results for positive INTONPQs (the left top and bottom facets) show a similar pattern in GigaAM CTC, GigaAM RNNT, and Whisper. Performance for the enriched audio stimuli with contexts related to question asking (the top left facet) is generally stable and, in some cases, slightly improves over the plain condition, although the gains are rather limited (only GigaAM RNNT does not exceed $60\,\%$ in Pre 'Next Q'). In contrast, the bottom left part of the plots, corresponding to conditions unrelated to questioning, shows a consistent decline, often reaching or falling below the plain baseline. This decrease is particularly visible for Whisper and GigaAM RNNT, where accuracy drops substantially compared to their baseline performance, while GigaAM CTC remains comparatively more robust, with a notable exception in the Pre R&L condition ($87.5\,\%$). The bootstrap results (provided in Table 2 in the Appendix) support the patterns observed in Figure 5, confirming that improvements for positive PQs are primarily associated with question-related contexts, while unrelated contexts tend to yield no effects or worsen detection. For negated PQs, Whisper shows better performance as well, but for the GigaAM models, the effect of context is minimal, with performance remaining largely stable across conditions.

Notably, there is a decrease in performance for Pre 'He asked' in LIPQs. In addition to matrix polar questions, the particle *li* also occurs in embedded contexts such as (5), resembling English *whether* or German *ob* (Restan, 1972; King, 1994).

(5) Mama sprosila, kupil li Maks xleb.
Mom asked bought LI Maks bread
'Mom asked whether Max bought bread.'

The drop is particularly visible for GigaAM RNNT: when the prepended context 'He asked' is combined with the presence of the *li* particle, the model tends to interpret the utterance as an embedded clause, whereas previously it placed a question mark. Although this is a possible interpretation, in our set it was still counted as incorrect, as two sentences were produced by a female and a male, meaning the models failed to recognize two different speakers.

4. Discussion

As rightly noted by the two anonymous reviewers, the experiment with Meta's MMS has a methodological limitation: the model itself does not predict punctuation but relies on an additional Silero component, which has no access to prosody. However, I decided to retain these results for three reasons. First, they point to a clear direction for improvement, namely integrating prosody and punctuation prediction. Second, MMS performs consistently poorly across conditions; i.e., it is stable in its errors, which makes its behavior interpretable rather than noisy. Third, this case also highlights a limitation of WER: while MMS may achieve a relatively good WER score, this metric does not capture its failure to detect sentence type.

When it comes to GigaAM and Whisper, the improved transcriptions observed when question-related utterances were added can plausibly be attributed to the architectures of the models. Whisper employs a transformer encoder–decoder model, while GigaAM is based on a Conformer architecture with self-attention. In both cases, attention mechanisms allow the models to make use of broader contextual information, which appears to improve question detection.

While this is encouraging, it also highlights a limitation of current ASR systems for Russian PQs. As is clear from above, prosody alone often signals sentence type and is reliably detected by native speakers without relying on additional contextual cues. At the same time, as rightly pointed out by an anonymous reviewer, human performance may also depend on factors such as language proficiency or sensitivity to contextual cues. However, all participants in the present study were native speakers of Russian, which minimizes the variability of the former issue. It is possible that non-native speakers would show lower accuracy or rely more on contextual information. For the latter issue, it is also conceivable that enriched audio could mislead human participants; for instance, by biasing them to interpret assertions as questions when preceded by cues such as *He asked* or *The next sentence is a question*. However, even in such cases, the effect would arise at the level of pragmatic expectations, rather than from an inability to process prosodic structure. By contrast, the ASR systems appear to rely on such external pragmatic cues in order to approximate distinctions that human listeners can derive directly from the acoustic signal.

Moreover, the models behave inconsistently across different contextual manipulations, which points to instability in their performance. This variability is not captured by standard WER, highlighting the need for more fine-grained evaluation measures sensitive to sentence type interpretation. While future work could extend the comparison to proprietary ASR systems or explore parameter settings such as temperature and beam size, it should no longer be treated as low-resource. Taken together, these results suggest that, for Russian, current ASR systems still fall short of human listeners in their ability to interpret sentence type from prosody alone.

Another crucial result from the experiment is that negation noticeably improves INTONPQs detection for multilingual Whisper and Russian-only GigaAM (accuracy is never below $60\,\%$), which, in turn, points to the fact that the models do rely on prosody somehow. Negative PQs might occur relatively often in the training data for Russian, so the models must have picked the combination of the nuclear pitch on the verb and negation as a question marker. I suggest that it is the combination of the two because (i) the models struggle with positive INTONPQs, i.e., with just verbal prominence, but (ii) for negative assertions, i.e., with no prominence on the verb, they perform with $100\,\%$ accuracy. It is unexpected because, cross-linguistically, negative PQs are considered to be marked (or biased; see, e.g., Gärtner and Gyuris 2017; Goodhue 2022) and used in specific contexts; thus, they should occur much less frequently than positive ones (for English, see Keisanen 2006). This also contradicts the results for Russian from the spoken corpus (Onoeva and Staňková, 2025): out of 500 randomly collected PQs, only 79 were negative ($15.8\,\%$). On the other hand, the so-called expletive negation is widely attested in Russian PQs (Brown and Franks, 1995; Abels, 2005; Zanon, 2024). Semantically, it is interpreted as having no negative force, but it might contribute a different meaning flavor in PQs, e.g., the inquirer's attitude towards a possible answer (see a similar idea for Czech in Šimík to appear). Mills (1992) brings further support for that claim, as she links negation in Russian PQs to politeness.

From an applied perspective, these findings raise questions about how current ASR systems handle prosodic information in downstream applications,

especially in tasks where sentence type is relevant, such as dialog systems or speech interfaces. At the same time, the present experiments show how controlled manipulations of input can be used to examine model behavior more systematically, in particular, their sensitivity to contextual cues.

5. Conclusion

In this article, I pursued the goal of going beyond WER as the primary metric in ASR evaluation and gaining insight into the "black box" behavior of state-of-the-art models. Punctuation proved to be a useful proxy in low-resource settings, allowing assessment of sentence type distinctions. The results suggest that standard WER-based evaluation would be insufficient here, as it does not capture systematic errors in prosodic interpretation. The experiments further indicate that current ASR systems rely on contextual cues and struggle to generalize from prosody alone, as reflected in their variability across conditions and the asymmetry between positive and negative INTONPQs. At the same time, their relatively strong performance on negative INTONPQs suggests that they do have access to prosodic information. However, reducing speech to text still overlooks other prosodic features, such as focus marking. Overall, this approach makes it possible to probe model behavior in a more fine-grained way.

6. Acknowledgments

I wish to express my gratitude to my co-authors of Razguliaeva et al. (to appear): Mariia Razguliaeva, Radek Šimík, Roland Meyer, and Kateřina Hrdinková. I also thank the organizers and committee of the SPEAKABLE workshop and the anonymous reviewers for their comments that improved the article. The study was funded by a one year grant from DAAD and CRC 1412 *Register* fellowship awarded to me in 2025–2026.

7. Data and code availability

Data and code are available on OSF: https://osf.io/4vcya/overview?view_only=d5bdad3e77d04e89981300236f634fbf

8. Ethics

Ethical approval was received for the QueSlav project (funded jointly by the Czech Science Foundation and the German Research Foundation). Human participants were reimbursed for the forced-choice task with €3.20. One participant was removed because they did not pass a reliability test (detect RAZVEPQs as PQs in 95 %).

9. Bibliographical References

Klaus Abels. 2005. "Expletive Negation" in Russian: A Conspiracy Theory. *Journal of Slavic Linguistics*, 13(1):5–74.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st workshop on benchmarking: Past, present and future*, pages 22–34.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (Version 5.1.13).

Sue Brown and Steven Franks. 1995. Asymmetries in the Scope of Russian Negation. *Journal of Slavic Linguistics*, 3(2):239–287.

Elena A. Bryzgunova. 1975. The declarative-interrogative opposition in russian. *The Slavic and East European Journal*, 19(2):155.

Matthew S. Dryer. 2013. Polar Questions (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Maria Esipova. 2025. Prosody across sentence types. *Semantics and Linguistic Theory*, pages 68–87.

Maria Esipova and Maribel Romero. 2023. Prejacent truth in rhetorical questions: Lessons from Russian. Talk at *Formal Approaches to Slavic Linguistics 32*.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2022. ESB: A Benchmark For Multi-Domain End-to-End Speech Recognition.

Hans-Martin Gärtner and Beáta Gyuris. 2017. On delimiting the space of bias profiles for polar interrogatives. *Linguistische Berichte*, 251:293–316.

Ljudmila Geist and Sophie Repp. 2023. Responding to negative biased questions in Russian. In Petr Biskup, Marcel Börner, Olav Mueller-Reichau, and Iuliia Shcherbina, editors, *Advances in formal Slavic linguistics 2021*, Open Slavic Linguistics. Language Science Press, Berlin.

Daniel Goodhue. 2022. Isn't there more than one way to bias a polar question? *Natural Language Semantics*, 30.

Lucas Rafael Stefanel Gris, Ricardo Marcacini, Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, and Sandra Maria Aluísio. 2023.

Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of life histories of the Museum of the Person.

Tiina Keisanen. 2006. *Patterns of stance taking: Negative yes/no interrogatives and tag questions in American English conversation*. @Acta Universitatis Ouluensis. Oulun Yliopisto, Oulu.

Tracy Holloway King. 1994. Focus in Russian Yes-No Questions. *Journal of Slavic Linguistics*, 2(1):92–120.

Natasha Korotkova. 2023. Conversational dynamics of Russian questions with *razve*. In *Proceedings of Sinn und Bedeutung 27*, Prague. Institute of Czech Language & Linguistic Theory, Faculty of Arts, Charles University.

Natasha Korotkova. submitted. A new perspective on negative bias in polar questions: The view from Russian. In G. Walkden Eckardt, R. and N. Dehé, editors, *The Oxford Handbook of Non-Canonical Questions*.

Aleksandr Kutsakov, Alexandr Maximenko, Georgii Gospodinov, Pavel Bogomolov, and Fyodor Minkin. 2025. GigaAM: Efficient Self-Supervised Learner for Speech Recognition. In *Interspeech 2025*, pages 1213–1217.

D. Robert Ladd. 2008. *Intonational Phonology*. Cambridge University Press.

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2023. LibriSpeech-PC: Benchmark for Evaluation of Punctuation and Capitalization Capabilities of End-to-End ASR Models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Roland Meyer. 2004. Prosody, Mood, and Focus. A Study of so-called "intonationally marked" Yes-No Questions in Russian. *Formal Approaches to Slavic Linguistics 12*, pages 333–352. The Ottawa Meeting.

Roland Meyer and Ina Mleinek. 2006. How prosody signals force and focus—A study of pitch accents in Russian yes–no questions. *Journal of Pragmatics*, 38(10):1615–1635.

Margaret H. Mills. 1992. Conventionalized politeness in Russian requests: A pragmatic view of indirectness. *Russian Linguistics*, 16(1).

Maria Onoeva and Anna Staňková. 2025. Polar questions in Czech and Russian: An exploratory corpus investigation.

OpenAI. 2022. Whisper. `https://github.com/openai/whisper`. GitHub repository, accessed February 2026.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Mariia Razguliaeva, Maria Onoeva, Radek Šimík, Roland Meyer, and Kateřina Hrdinková. to appear. Processing of Russian and Czech polar questions: Evidence for the effect of question bias. *Glossa Psycholinguistics*.

Steve Renals and Simon King. 2010. Automatic Speech Recognition.

Per Restan. 1972. *Sintaksis voprositel'nogo predloženija: obščij vopros [Syntax of interrogative sentences: polar questions]*. Universitetsforlaget, Oslo.

Salute Developers. 2023. GigaAM: Large-Scale Russian Automatic Speech Recognition. `https://github.com/salute-developers/GigaAM`. GitHub repository, accessed February 2026.

Silero Team. 2026. Silero Models: Pre-trained enterprise-grade STT/TTS models and benchmarks. GitHub repository, accessed February 2026.

Radek Šimík. to appear. *Polar question semantics and bias: Lessons from Slavic/Czech*.

Ksenia Zanon. 2024. Expletive Negation revisited: on some properties of negative polar interrogatives in Russian. *Journal of Slavic Linguistics*.

10. Appendices

Table 2: Performance across the ASR models (excluding Meta's MMS) for INTONPQs only under different contextual conditions. Accuracy (Acc.), improvement (Imp.) over baseline, and 95 % bootstrap confidence intervals (CI) are reported separately for positive and negative INTONPQs. Improvements are expressed in percentage points relative to the baseline (plain audio) condition from Table 1. Effects are classified as significant when the confidence interval does not include zero. The rest of the results are available on OSF.

Context	Positive INTONPQs				Negative INTONPQs			
	Acc.	Imp.	CI	Effect	Acc.	Imp.	CI	Effect
Whisper								
Pre 'Next Q'	78.10	71.90	[56.25, 87.50]	Improves	100.00	37.50	[21.88, 53.12]	Improves
Pre 'He asked'	68.80	62.50	[43.75, 78.12]	Improves	100.00	37.50	[21.88, 53.12]	Improves
Post 'yes'	90.60	84.40	[71.88, 96.88]	Improves	100.00	37.50	[21.88, 56.25]	Improves
Post 'no'	87.50	81.20	[65.62, 93.75]	Improves	100.00	37.50	[21.88, 53.12]	Improves
Pre 'Turtle'	12.50	6.20	[−6.25, 18.75]	No effect	96.90	34.40	[18.75, 50.00]	Improves
Post 'Turtle'	15.60	9.40	[−3.12, 21.88]	No effect	75.00	12.50	[3.12, 25.00]	Improves
Pre R&L	12.50	6.20	[−6.25, 18.75]	No effect	90.60	28.10	[12.50, 43.75]	Improves
Post R&L	3.10	−3.10	[−9.38, 0.00]	No effect	71.90	9.40	[−6.25, 25.00]	No effect
Pre 'He said'	25.00	18.80	[6.25, 34.38]	Improves	100.00	37.50	[21.88, 53.12]	Improves
GigaAM: RNNT								
Pre 'Next Q'	59.40	25.00	[3.12, 46.88]	Improves	100.00	0.00	[0.00, 0.00]	No effect
Pre 'He asked'	96.90	62.50	[46.88, 78.12]	Improves	100.00	0.00	[0.00, 0.00]	No effect
Post 'yes'	96.90	62.50	[46.88, 78.12]	Improves	96.90	−3.10	[−9.38, 0.00]	No effect
Post 'no'	96.90	62.50	[46.88, 78.12]	Improves	96.90	−3.10	[−9.38, 0.00]	No effect
Pre 'Turtle'	3.10	−31.20	[−46.88, −15.62]	Worsens	84.40	−15.60	[−28.12, −3.12]	Worsens
Post 'Turtle'	18.80	−15.60	[−31.25, 0.00]	No effect	87.50	−12.50	[−25.00, −3.12]	Worsens
Pre R&L	28.10	−6.20	[−28.12, 15.62]	No effect	100.00	0.00	[0.00, 0.00]	No effect
Post R&L	28.10	−6.20	[−25.00, 12.50]	No effect	93.80	−6.20	[−15.62, 0.00]	No effect
Pre 'He said'	6.20	−28.10	[−43.75, −12.50]	Worsens	65.60	−34.40	[−50.00, −18.75]	Worsens
GigaAM: CTC								
Pre 'Next Q'	87.50	31.20	[12.50, 50.00]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Pre 'He asked'	100.00	43.80	[28.12, 59.38]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Post 'yes'	100.00	43.80	[28.12, 62.50]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Post 'no'	100.00	43.80	[28.12, 59.38]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Pre 'Turtle'	25.00	−31.20	[−46.88, −15.62]	Worsens	93.80	−3.10	[−9.38, 0.00]	No effect
Post 'Turtle'	53.10	−3.10	[−18.75, 12.50]	No effect	96.90	0.00	[0.00, 0.00]	No effect
Pre R&L	87.50	31.20	[15.62, 46.88]	Improves	100.00	3.10	[0.00, 9.38]	No effect
Post R&L	50.00	−6.20	[−21.88, 9.38]	No effect	96.90	0.00	[0.00, 0.00]	No effect
Pre 'He said'	18.80	−37.50	[−53.12, −21.88]	Worsens	87.50	−9.40	[−21.88, 0.00]	No effect