Disfluencies and ASR Performance on Swedish Spontaneous Speech from the 'Trip to Stockholm' Discourse Narrative Task

Dimitrios Kokkinakis, Herbert Lange, Ricardo Muñoz Sánchez
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg, Sweden
{dimitrios.kokkinakis, herbert.lange, ricardo.munoz.sanchez}@svenska.gu.se

Abstract

Automatic Speech Recognition (ASR) offers a scalable and cost-efficient alternative to manual transcription and is becoming increasingly relevant in clinical contexts, particularly for the detection of cognitive decline and mental health assessment. However, current ASR-systems still struggle with spontaneous speech, particularly when processing disfluencies, pauses, and speaker variability that often carry diagnostic value. This study evaluates state-of-the-art open ASR models targeting Swedish using recordings from the "Trip to Stockholm" discourse narrative task which elicits ecologically valid, cognitively demanding speech. Recognition quality is assessed using various metrics, alongside an analysis of linguistic and technical sources of error focused on disfluencies. Our findings show that disfluency-related phenomena degrade recognition performance. Possible post-processing strategies can improve specific error patterns emerging for filled pauses, word repetitions, and self-corrections. The results illustrate both the advances *and* ongoing limitations of ASR for spontaneous Swedish speech, emphasizing the need for models explicitly trained, or fine-tuned, on disfluent data to ensure robustness in clinical and research applications.

Keywords: Whisper, KB-Whisper, speech-to-text, disfluencies, discourse task, cognitive impairment, Swedish

1. Introduction

Automatic Speech Recognition (ASR) is a scalable and cost-efficient technology that addresses the increasing volume of digital content and the corresponding demand for accessible communication. ASR offers several advantages, including low deployment costs, minimal logistical and availability constraints, and the ability to operate continuously without interruption, making it particularly attractive for clinical applications such as mental health assessment and the detection of cognitive decline. An important advantage of ASR is its potential to mitigate the cost and time demands of manual transcription while reducing transcription errors and forms of human-induced bias, such as subjective interpretation and inconsistencies across transcribers. Given the privacy and recording constraints frequently encountered in clinical settings, ASR provides a scalable means of capturing spoken responses for subsequent analysis.

However, ASR errors—particularly failures to accurately capture disfluencies and pauses that are central to cognitive assessment—constitute a persistent and consequential limitation, potentially obscuring dementia-specific speech markers (cf. Li et al., 2024). In this study, we evaluate the performance of automatic speech recognition systems on Swedish speech data, with a specific focus on disfluencies—excluding pauses—by assessing recognition outcomes using Word Error Rate (WER) and complementary metrics. We examine linguistic factors and technical constraints underlying transcription errors and systematically analyze common error patterns in spontaneous speech across different hyperparameter settings. Since performance metrics quantify the deviation of a transcript from a reference, a central aim of this study is to also discuss ways to improve transcription output.

We apply the performance metrics to a realistic dataset by first applying state-of-the-art open ASR models and assessing their ability to handle diverse linguistic structures and speech variability across varying speaker groups. Recordings from the "Trip to Stockholm" task were used as input for evaluating ASR models. This is a spoken discourse task which was modeled after the "Trip to New York" task described in Harris et al. (2008). The spontaneous and variable nature of these speech samples provides a valuable benchmark for evaluating ASR performance in ecologically valid, cognitively demanding, Swedish discourse settings.

2. Background and Related Work

Speech disfluencies — such as repetitions, self-corrections, filled pauses and other interruptions in the flow of speech — are a natural feature of spontaneous language and have been widely studied for their informational value. Additionally, disfluencies are well-established indicators of cognitive and mental health status and have been widely associated with neurodegenerative conditions such as dementia, particularly Alzheimer's disease (AD).

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 24–33
12 May 2026.

Such non-fluent speech patterns have been shown to increase with disruptions in language planning, executive functioning, and cognitive load (Jiang and An, 2025; Clark and Fox Tree, 2002), highlighting how such phenomena relate to cognitive and lexical-semantic impairment in AD (Pistono et al., 2024).

State-of-the-art ASR systems do not adequately capture and label word- and phrase-level disfluencies (Shahla et al., 2022; Cumbal et al., 2024). In addition, in a diachronic context, age correlates with changes in speech rate and lexical complexity, which, in turn, contribute to increased production of disfluency over time, establishing disfluency as a robust marker of age-related linguistic change (Beier et al., 2023). As Nasreen et al. (2021) demonstrated, disfluency features in conversational speech serve as noninvasive biomarkers of moderate-stage Alzheimer's disease, revealing significant differences between AD and age-matched non-AD participants. Modern ASR systems often treat disfluencies as noise and remove them during post-processing,[1] obscuring potentially meaningful details in the transcript (Dinkar, 2022).

During the last couple of years, there have been notable advances in automatic speech recognition for Scandinavian languages, for example Norwegian,[2] and, more importantly for our study, ASR models trained on Swedish data are now readily available (Vesterbacka et al., 2025; Li et al., 2025) – see Section 3.4.

However, such models continue to exhibit substantial performance gaps across a range of neuropsychological assessment settings, largely because they are not trained on acoustically challenging recording conditions, nor on stylistic variability in naturalistic speech production, or speaker characteristics commonly encountered in mental health contexts and early cognitive decline or dialects. Kokkinakis et al. (2025) reported a much higher WER (0.265) on a picture description task ("Cookie Theft") and a substantially lower WER (0.033) on a reading-aloud task, which imposes fewer demands on spontaneous speech production. Such results indicate that ASR systems *can* achieve near-human transcription accuracy on controlled reading tasks, while more spontaneous, cognitively demanding speech—such as narrative or descriptive tasks—often result in higher error rates and reduced reliability for downstream analyses.

[1] Short words may be omitted for a variety of reasons, including insufficient or noisy acoustic evidence, the merging or normalization of tokens during transcription, and the tendency of language models to favor paraphrased outputs that exclude short function words when these result in more probable overall sequences.

[2] Available from: https://huggingface.co/NbAiLab/nb-whisper-small-beta.

3. Participants, Dataset, Format and Technical Details

Originally, the subset of the participants used in the current study were recruited from the *Gothenburg MCI study*, a longitudinal study investigating dementia disorders in patients seeking medical care at a memory clinic (Wallin et al., 2016). The data analyzed here consist of recordings of spontaneous speech collected as part of the separate research project "Linguistic and extra-linguistic parameters for early detection of cognitive impairment" (Riksbankens Jubileumsfond, NHS 14-1761:1 - 2016-2020).

The audio recordings were collected in a relatively controlled environment and feature native Swedish speakers as their first language. The study was approved by ethical approval (reference number: 206–16, 2016; T021-18) issued by the regional ethical review board in Gothenburg, Sweden. Participants were informed that they could withdraw their participation at any time. All data were coded and anonymized. For the audio capture of the task, a Zoom H4n Handy recorder was used, and the resulting audio files were saved and stored as uncompressed audio in .wav-format 44.1 kHz with 16-bit resolution. A speech pathologist and computational linguist were present during the recording sessions, providing all subjects with identical instructions according to a predefined protocol (cf. Section 3.2). The audio recordings were manually transcribed by a professional transcription company using a standardized procedure and clearly specified guidelines to ensure consistency and accuracy. Manual transcribers added full stops at the end of sentences because the same data should be usable for syntactic parsing, which requires explicit sentence boundary marking. All data were stored within a secure university-managed infrastructure, using a information security class 3 platform based on Nextcloud,[3] in compliance with institutional data protection requirements.

3.1. Demographic and Dataset Characteristics

The study sample consisted of speech recordings from 30 participants in the aforementioned study, drawn from *Västra Götaland County* in Sweden. The participants ranged in age from 57 to 78 years (M = 68.9, SD $\approx$ 5.37). The sample included an equal percentage of female/male participants and also equally distributed across the three categories:

[3] https://nextcloud.com.

	$HC_{n=10}$	$MCI_{n=10}$	$SCI_{n=10}$
Females	5	5	5
Males	5	5	5
Mean Age	71.2	68,4	67
Mean Years of Education	13.5	12.8	16.3
Mean Recording Duration	158.3s	130.8s	194s
Mean Token Number	402.5	294,3	426.4
Number of Tokens with Disfluencies	93	86	61
Type-Token Ratio	4.30	3.83	3.82

Table 1: Demographic information for the three groups of participants.

Healthy Controls, HC,[4] *Mild Cognitive Impairment, MCI,*[5] and *Subjective Cognitive Impairment, SCI.*[6] Details of demographic information for the three groups of participants are shown in Table 1.

In the table, Type–Token Ratio (TTR) is a measure of lexical diversity, defined as the number of unique words (types) divided by the total number of words (tokens) in a text or speech sample. In our sample, the TTR is between 4.30 (higher for the healthy group) and 3.82, which is typical for spontaneous informal speech. A higher TTR implies greater lexical diversity, which is often desirable in speech. However, no statistical significance claims can be made due to the small sample size, and the measurements in Table 1 are just given as a reference.

3.2. "Trip to Stockholm": a Swedish Spoken Discourse Task for ASR Evaluation

The spontaneous language material analyzed in the present study was derived from a spoken discourse task modeled on the "Trip to New York" (Harris et al., 2008; Fleming and Harris, 2008). For the purposes of this project, the task was adapted to "Trip to Stockholm" (Antonsson et al., 2021). The use of this complex discourse task has shown the potential to differentiate adults who are normally aging cognitively from those with MCI (Fleming, 2014), which is an important task of our research. Participants were asked to describe orally how they would plan and carry out a trip to Stockholm, following a short series of instructions about the planning of a two week trip:

> *Now you are going to do a task where you are asked to think and plan aloud. Imagine that you are going on a vacation a week from now. You are traveling to Stockholm for a 2-week stay. Think about all you will have to do to get ready to go, such as how you will get there, what you will bring, and what you will do. I want you to tell me all of your plans until I ask you to stop after about 4 to 5 min.*

If participants did not spontaneously include certain information in their narratives, brief follow-up prompts were provided (e.g., "Who will take care of your mail?" or "What will you bring on your trip?"). The task was designed to elicit connected, naturalistic speech requiring conceptual and semantic elaboration related to the cognitive-linguistic schema for travel (Harris et al., 2008). Due to its cognitive and linguistic complexity, the task has been suggested to be sensitive to subtle deficits in individuals with brain injury, engaging executive functions such as initiation, planning, temporal organization, and flexibility, as well as semantic, episodic, and working memory processes.

3.3. Pre-processing

Common disfluency features in Swedish, such as *nonlexical fillers* ”hm”, ”eh” and *vocalizations* "haha", as well as *false starts*, were *not* omitted. Word fragments, e.g., *‘[...] inte betal- beställa något hotell [...]’* (lit. [...] not *pay-* to book a hotel [...]) were transcribed as complete words whenever the intended word could be reliably identified by the manual transcriber; if not, the transcription preserved the original partial or interrupted form; as in *‘[...] när jag har txxx jag tycker [...]’* (lit. [...]’ when I have *txxx*, I think [...]).

Numerical data as well as occurrences of URLs, were rendered in full; for example, "E4"[7] was transcribed as *e-four*; and *bookings.com* as three tokens *bookings punct com*.

In the evaluation all tokens were converted to lowercase and punctuation marks were removed. The Python package `werpy` was used for text normalization.[8] The aim of these normalizations was to

[4] *Healthy Controls* are participants who do not exhibit the condition under investigation and serve as a baseline or comparison group against affected individuals.

[5] *Mild Cognitive Impairment* is defined as a transitional stage of cognitive decline that lies between the alterations associated with normal aging and the deficits that satisfy the diagnostic criteria for clinical dementia (Petersen et al., 2014; Albert et al., 2011).

[6] *Subjective Cognitive Impairment* refers to a self-perceived, persistent decline in one or more cognitive domains over time, occurring in the absence of objectively measurable deficits (Jessen et al., 2007).

[7] "E4" is a major European route (motorway/highway).

[8] https://pypi.org/project/werpy/.

improve the accuracy when matching the transcription with the gold data since difference in phrase segmentation and inconsistencies in using upper-/lowercase letters can have detrimental effects on the evaluation.

3.4. Models and Metrics

Most modern ASR systems are based on OpenAI's Whisper (Radford et al., 2022; OpenAI, 2022), which uses a sequence-to-sequence transformer architecture. Audio is converted to a log-Mel spectrogram, encoded, and then decoded autoregressively into text tokens (words, subwords, or punctuation). Processing occurs in independent segments, which are later combined, allowing efficient transcription but occasionally producing local inconsistencies.

In the present study, ASR transcriptions were generated using locally deployed versions of three publicly available models and variants:

- OpenAI Whisper is an ASR model trained in more than 680,000 hours of multilingual, multitask audio data, designed to support robust transcription and translation across a wide range of languages and recording environments (Radford et al., 2022).[9]
- the Swedish National Library's KB-Whisper, is based on OpenAI's Whisper architecture but trained on over 50,000 hours of Swedish audio (Vesterbacka et al., 2025).[10] The training corpus included TV broadcasts, parliamentary debates, and dialectal recordings, yielding substantially improved accuracy for Swedish speech compared to the original OpenAI model.
- Stable-TS, a variant of OpenAI Whisper, is an open-source timestamp refinement and alignment layer for Whisper-based ASR. It post-processes OpenAI Whisper model output by re-aligning text tokens to audio using forced-alignment and smoothing heuristics.[11]

All three systems were used in their `Faster-Whisper` implementation[12] which utilizes the `CTranslate2` library,[13] a fast inference engine for Transformer models.

We used four size versions of each model: *tiny*, *small*, *medium*, and *large*. Each model configuration, model type, and size, was also assessed using three primary evaluation metrics:

- Word Error Rate (WER): which quantifies the proportion of words incorrectly predicted by a model, accounting for substitutions, deletions, and insertions relative to the reference transcription; that is, the minimum edit distance between a transcript and the reference (ground truth), expressing the proportion of errors relative to the total number of words. The WER metric typically ranges from 0 to 1, where 0 indicates that the compared pieces of text are exactly identical, and 1 (or larger) indicates that they are completely different with no similarity. A WER of 0.8 means that there is an error rate of 80% for the compared sentences.
- Bilingual Evaluation Understudy (BLEU): which measures the n-gram overlap between the predicted and reference transcriptions; (Papineni et al., 2002). BLEU computes a value between 0 and 1, where 1 corresponds to perfect agreement between the prediction and the gold standard. Although BLEU was originally developed for machine translation evaluation, it has also been applied to speech-to-text output by comparing the generated transcript with a reference text. However, this metric does not fully capture recognition errors and should therefore be interpreted with caution.
- Google-BLEU (GLEU): a measure intended to overcome limitations in BLEU score calculations and are better suited for sentence level comparisons GLEU balances precision and recall over 1-4 n-grams between predicted and reference transcriptions. (Mutton et al., 2007).

We used the BLEU and GLEU implementations from NLTK[14] as well as the WER implementation from the `werpy` package.

The study also considers a range of the hyperparameter *temperature* settings (0, 0.25, 0.5, 0.75, and 1), which control the degree of randomness during decoding. Lower temperature values lead to more deterministic and stable transcriptions, whereas higher values introduce increased variability in the generated output, potentially capturing alternative word choices at the cost of consistency, i.e. increased error rates (cf. Table 2).

[9] https://huggingface.co/openai/whisper-large-v3.
[10] https://huggingface.co/KBLab/kb-whisper-large.
[11] https://github.com/jianfch/stable-ts.
[12] https://github.com/SYSTRAN/faster-whisper.
[13] https://github.com/OpenNMT/CTranslate2/.
[14] https://www.nltk.org/.

4. Evaluation and Analysis

The three ASR models are assessed against the reference transcripts, using a version without any punctuation markings. The transcripts were pre-processed according to the previous description (Section 3.3). Numerical tokens were converted to text (e.g. "4" to "four") and all transcriptions were transformed to lower-case. We evaluated the performance of each ASR model using the previously described metrics, WER, BLEU, and GLEU, and the five temperature hyperparameters.

Swedish short discourse adverbs and function words (2-3 characters long), such as *väl* (lit. “well”) and *ju* (a discourse particle roughly meaning “as you know” or “after all”) are often dropped in transcriptions, probably because of acoustic ambiguity and language model bias. For example, *ska väl gå* becomes *ska gå* (lit. “it should work”) or *’a så har de ju vasamuseet’* becomes *’och så har de vasamuseet’* (lit. “and then they have the Vasa Museum, of course”); see also footnote 1. Similar behavior is observed in multiword function words such as *i och med* (“given that” or “due to”); *nu är jag ju bortskämd i och med att jag har en hustru* (lit. “I am, of course, spoiled, given that I have a wife”) to *nu är jag bortskön att jag har en hustru* (lit. “Now I am ‘bortskön’ that I have a wife”); note also the wrong annotation of *bortskämd* to *bortskön*.

Some other discrepancies between near-verbatim manual transcription and the models’ output, as well as between orthographic and phonetic-near transcriptions can be explained by:

(i) homonym-phonetic confusion (’å’ sometimes ’och’ [and]);

(ii) occasional cases in which the orthographic transcription incorporated manually asserted phonetic symbols (e.g., ’n:u’ [now] and ’- -’ longer pauses), reflecting a partial convergence with phonetic-level representation;

(iii) phonological deviation ’å slå sej’ (lit. ’och slå sig’) – [and beat themselves].

More importantly, *ASR artefacts*, such as the outputted word “lagoda”, may result from word concatenation, erroneous normalization, or phonetic approximation under conditions of rapid or indistinct Swedish speech; in this instance, the form most likely corresponds to a misrecognition of the proper name “Agoda”, a travel agency.

Finally, *overregularization*, when a grammatically valid rule is applied in an inappropriate linguistic context, is also observed in such cases models apply the regular Swedish plural suffix (here ’-ar’) to nouns that exhibit zero plural marking in standard Swedish; e.g. ’fyra lamm’ (lit. four lambs) to ’fyra lammar’; or ’skjortor’ (lit. shirts) as ’skjortar’.

4.1. Performance Results

The evaluation results are shown in Table 2. As can be seen in this table there are systematic differences in performance depending both on the type of model and the size configuration. Larger configurations generally yield higher accuracy, though gains are model-dependent, and some smaller configurations achieve competitive results, suggesting potential trade-offs between computational cost and performance.

In all tasks, as expected, the large Swedish *KB-Whisper* model performed overall best for all three metrics, with a temperature set to 0.5. In fact, 0.147 was the best overall WER value, 0.927 the best overall value for BLEU, and 0.926 the best overall GLEU value, regardless of the model. Lower WER values generally indicate transcripts that more closely approximate the reference text and are therefore more likely to be understandable. While high BLEU and GLEU values indicate that the generated transcripts closely match the reference text in terms of word choice and local phrase structure. Notably, the lowest WER values for *OpenAI* and *Stable-TS* were the medium models with temperature=0, 0.202 and 0.199, respectively. As an outlier the *OpenAI-large* model started repeating a part of the output for one of the example over and over again which resulted in very bad scores. We are not sure about the source of this problem but were able to reproduce with the same input and parameters.

The results show consistent differences in the processing/transcription time between model sizes and resource settings.[15] For both KB-Whisper and OpenAI Whisper, the *large* models require substantially longer total runtime than the *tiny* variants (approximately 13–14 minutes vs. just over 8 minutes), with average per-instance processing times of around 5.5 seconds for *large* and 3.3 seconds for *tiny*. In contrast, the Stable-TS setting is markedly more computationally demanding, with runtimes increasing by a factor of approximately four for all models, particularly for the *large* configuration (56 minutes total, 22.5 seconds on average).[16] Despite these differences, the relative efficiency gap between *large* and *tiny* remains stable across settings, indicating that model size consistently impacts computational cost, while the choice of resource configuration has a much stronger effect on overall runtime.

[15] All data has been transcribed using the CUDA implementation of Faster Whisper on a server equipped with a NVIDIA GeForce RTX 3060 (12 GB RAM).

[16] The length of the audio recordings is ca 80 min, all transcriptions were substantially faster than real-time.

			temp=0	temp=0.25	temp=0.5	temp=0.75	temp=1
KB-Whisper	large	WER	0.149	0.153	**0.147***	0.163	0.189
		BLEU	0.926	0.922	**0.927***	0.916	0.903
		GLEU	0.924	0.920	**0.926***	0.914	0.900
	medium	WER	**0.206**	**0.206**	0.207	0.216	0.219
		BLEU	**0.879**	0.881	0.877	0.869	0.870
		GLEU	0.874	**0.876**	0.872	0.863	0.864
	small	WER	**0.171**	0.173	0.179	0.182	0.196
		BLEU	**0.913**	0.909	0.910	0.903	0.895
		GLEU	**0.912**	0.907	0.908	0.901	0.891
	tiny	WER	0.565	0.683	0.583	**0.244**	0.279
		BLEU	0.571	0.499	0.534	**0.885**	0.867
		GLEU	0.505	0.415	0.447	**0.882**	0.863
OpenAI	large	WER	0.674	0.273	0.658	**0.224**	0.273
		BLEU	0.837	0.862	0.799	**0.875**	0.847
		GLEU	0.835	0.859	0.796	**0.869**	0.840
	medium	WER	**0.202***	0.218	0.227	0.245	0.364
		BLEU	**0.904***	0.893	0.883	0.868	0.790
		GLEU	**0.902***	0.890	0.878	0.863	0.776
	small	WER	**0.253**	0.290	0.298	0.317	0.476
		BLEU	**0.880**	0.849	0.851	0.841	0.729
		GLEU	**0.876**	0.842	0.845	0.834	0.708
	tiny	WER	**0.476**	0.646	0.584	0.605	0.774
		BLEU	**0.795**	0.727	0.759	0.726	0.544
		GLEU	**0.785**	0.716	0.748	0.706	0.491
Stable-TS	large	WER	0.220	0.274	0.310	**0.210**	0.268
		BLEU	**0.893**	0.867	0.855	0.886	0.855
		GLEU	**0.890**	0.865	0.852	0.882	0.849
	medium	WER	**0.199***	0.234	0.224	0.237	0.365
		BLEU	**0.904***	0.874	0.886	0.879	0.788
		GLEU	**0.902***	0.869	0.883	0.875	0.770
	small	WER	**0.259**	0.283	0.292	0.326	0.490
		BLEU	**0.874**	0.855	0.852	0.834	0.716
		GLEU	**0.870**	0.849	0.845	0.826	0.692
	tiny	WER	**0.510**	0.537	0.572	0.603	0.772
		BLEU	**0.770**	0.760	0.739	0.728	0.564
		GLEU	**0.759**	0.746	0.723	0.708	0.521

Table 2: Evaluation results for the three main models, *KB-Whisper*; *OpenAI*; *Stable-TS*, comparing performance across the four size configurations (*tiny*, *small*, *medium* and *large*) and five temperature values for each model architecture (*0*, *0.25*, *0.5*, *0.75* and *1*). The best results for each temperature value are marked in bold and the overall best result for each model is marked with an asterisk (*).

5. Discussion and Future Work

Despite major advances in ASR and claims of near-human precision, evaluations in domains such as Higher Education lectures reveal substantial variability and reduced reliability, particularly for streaming applications (Kuhn et al., 2024). Although our study is subject to limitations, such as the inclusion of only 30 participants, the findings may nonetheless provide valuable insights for downstream applications. In particular, when combined with careful preprocessing and quality control, these approaches can support automated cognitive evaluation and the monitoring of language-related decline, especially in large-scale evaluations involving population-level cohorts.

We strongly believe that the WER (and other metrics) on this kind of data can be improved through several practical strategies, including enhancing audio quality, employing domain-specific language models, applying post-processing corrections, or retraining the system with additional in-domain speech data. These approaches are primarily applicable in future data collection scenarios and can involve recordings of individuals with mild or severe cognitive impairments, although the collection of such data necessarily requires careful ethical consideration.

By contrast, we want to explore *prompting*. We can apply this method directly to existing recordings without the need of additional data acquisition or training. Further assessment of the robustness of this framework, together with evaluation of its performance, constitutes a primary focus of future work. Prompts for Whisper models are used to stitch together multiple audio segments, Whisper is using a sliding audio context window of 30 seconds.[17] However, giving an initial prompt can even steer the model output, providing spelling and output formatting hints.[18]

In addition, future investigations will extend the evaluation to newly released and updated Whisper-inspired model versions,[19] enabling a more comprehensive comparison.

On the research infrastructure side we work on extending the tooling provided by Språkbanken Text. Whisper-based transcriptions (Språkbanken Text, 2025c) are already available in the Sparv pipeline (Språkbanken Text, 2025d) and consequently also in the Mink platform (Språkbanken Text, 2025a,b). However, the feature set is currently too limited to conduct our experiment within this framework. We plan to add the missing features required by our experiments, allowing us as well as other researchers to reproducibly repeat this and similar experiments using the Sparv (Språkbanken Text, 2025d) pipeline. The full Whisper toolbox will be made available within Mink. More details about Mink and Sparv can be found in Forsberg et al. (2025). The current code for the experiment and evaluation is also available on Github under a free and open license.[20]

[17] https://github.com/openai/whisper.

[18] https://developers.openai.com/cookbook/examples/whisper_prompting_guide.

[19] Other models *not* considered in the study include https://huggingface.co/birgermoell/whisper-small-sv-bm and the https://github.com/m-bain/whisperX.

[20] https://github.com/spraakbanken/Whisper-experiment/.

6. Conclusion

This study evaluates state-of-the-art automatic speech recognition (ASR) models for the full-scale automatic transcription of a Swedish discourse narrative task. The population in focus includes individuals with early signs of cognitive impairment. ASR provides a scalable and automated method for analyzing spoken responses in cognitive assessments. However, spontaneous speech—often characterized by disfluencies, hesitations, and complex syntactic structures—remains more challenging than controlled reading tasks, influencing recognition accuracy and downstream analysis such as automatic scoring or classification tasks.

We focus on Word Error Rate (WER) because it directly measures word-level transcription accuracy, which is the primary objective of this work. WER reflects human correction effort, penalizes substitutions, deletions, and insertions symmetrically, and discourages over-generation and hallucination. Although WER does not capture semantic equivalence, it is a deliberate and widely accepted choice to evaluate transcription fidelity and ensures comparability with previous ASR literature. In addition to WER, we report metrics in BLEU and GLUE-style to capture complementary aspects of ASR output quality. WER measures exact word-level transcription fidelity, whereas BLEU reflects local phrase consistency and fluency by providing partial credit for benign lexical variations. GLUE-style metrics assess semantic equivalence, enabling us to distinguish meaning-preserving deviations from semantically harmful errors. Together, these metrics offer a more complete evaluation while retaining WER as the primary measure of transcription accuracy.

This evaluation therefore offers practical guidance for selecting ASR models suited to Swedish-language clinical and research applications, balancing transcription quality with robustness to natural speech variations. By mapping the performance of the model to the demands of specific tasks, we outline a framework for integrating AI transcription into screening workflows. The findings underscore the importance of task-sensitive model evaluation and support the development of automated tools and platforms for cognitive evaluation.

7. Limitations

The dataset is limited by the small sample size, comprising only 30 participants from the same geographical area and with a comparable age and level of education. This restricts the generalizability of the findings and calls for caution when interpreting the results.

AI models can produce confusing words (or sentences) by mistaking homonyms or hallucinating text, particularly in cases where contextual cues are weak and the model must rely on uncertain predictions (Koenecke et al., 2024). Moreover, evaluation metrics such as WER — though simple and easy to compute — have been criticised for not capturing text understanding and for correlating only weakly with human judgments of transcript quality (Just et al., 2025; Phukon et al., 2025). Still, WER remains one of the most widely used and practical metrics for evaluating ASR systems, as their verbatim outputs lend themselves to word-by-word comparison.

8. Ethical Considerations

During the experiments, we ensured that no private or personally identifiable information—such as participants' names and health data — was disclosed or processed outside the local environment. To minimize privacy risks and maintain full control over the data, all experiments were conducted exclusively using locally installed open-source models. This approach ensured that no data were transmitted to external servers or third-party services, thereby complying with data protection and ethical research standards.

9. Disclosure

The authors used the digital assistant platform *ChatGPT Edu version 5.2*, to support limited aspects of the writing process, specifically grammar, morphological refinement, and related checks (e.g., spelling verification and typographical correction). All conceptual contributions, analyses, interpretations, and conclusions are solely the authors' own, and no generative tool was used to produce or modify empirical data, figures, or results.

10. Acknowledgments

The research presented here was supported by the Swedish Research Council (grant number 2025-00765), the Swedish national research infrastructure Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (grant numbers 2017-00626 and 2023-00161), the Huminfra, the Swedish national infrastructure for the Humanities, funded by the Swedish Research Council and the consortium nodes (grant numbers 2021-00176 and 2023-00171), as well as by the The Swedish Parkinson Foundation (Parkinsonfonden).

11. Bibliographical References

Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. 2011. The diagnosis of mild cognitive impairment due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279.

Malin Antonsson, Kristina Lundholm Fors, Marie Eckerström, and Dimitrios Kokkinakis. 2021. Using a discourse task to explore semantic ability in persons with cognitive impairment. *Frontiers Aging Neuroscience*, 12.

Eleonora J Beier, Suphasiree Chantavarin, and Fernanda Ferreira. 2023. Do disfluencies increase with age? evidence from a sequential corpus study of disfluencies. *Psychology and Aging*, 38(3):203–218.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Ronald Cumbal, Birger Moëll, Jose Lopes, and Engwall Olof. 2024. "you don't understand me!": Comparing asr results for l1 and l2 speakers of swedish. *Computing Research Repository*, arXiv:2405.13379v1. Version 1.

Tanvi Dinkar. 2022. Computational models of disfluencies : fillers and discourse markers in spoken language understanding. *Computer science. Institut Polytechnique de Paris*, NNT: 2022IPPAT001.

Valarie Fleming and Joyce L. Harris. 2008. Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology*, 22:792–740.

Valarie B. Fleming. 2014. Early detection of cognitive-linguistic change associated with mild cognitive impairment. *Communication Disorders Quarterly*, 35:146–157.

Markus Forsberg, Dana Dannélls, Lars Borin, and Aleksandrs Berdicevskis. 2025. *Background: Språkbanken Text*, chapter 9. De Gruyter.

Joyce L. Harris, Swathi Kiran, Thomas P. Marquardt, and Valarie B. Fleming. 2008. Communication wellness check-up©: Age-related changes in communicative abilities. *Aphasiology*, 22:813–825.

Frank Jessen, Birgitt Wiese, Gabriela Cvetanovska, Angela Fuchs, Hanna Kaduszkiewicz, Heike Kölsch, Tobias Luck, Edelgard Mösch, Michael Pentzek, Steffi G Riedel-Heller, Jochen Werle, Siegfried Weyerer, Thomas Zimmermann, Wolfgang Maier, and Horst Bickel. 2007. Patterns of subjective memory impairment in the elderly: association with memory performance. *Psychol Med.*, 37(12):1753–62.

Yue Jiang and Xufei An. 2025. Speech differences between aged women with and without early alzheimer's disease: linguistic indicators of cognitive decline. *Acta Psychologica*, 256.

Sandra Anna Just, Brita Elvevåg, Shrankhla Pandey, Ivan Nenchev, Anna-Lena Bröcker, Christiane Montag, and Sarah E Morgan. 2025. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders. *Psychiatry Research*, 352.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681, New York, NY, USA. Association for Computing Machinery.

Dimitrios Kokkinakis, Herbert Lange, and Ricardo Muñoz Sánchez. 2025. Evaluating speech-to-text models for swedish neuropsychological assessments: a comparative study across task types and models. In *Proceedings of the 35th Alzheimer Europe conference "Connecting science and communities: The future of dementia care*, Bologna, Italy.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2024. Measuring the accuracy of automatic speech recognition solutions. *ACM Trans. Access. Comput.*, 16(4).

Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2024. Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics*, 150.

Zirui Li, Jens Edlund, Yicheng Gu, Nhan Phan, Lauri Juvela, and Mikko Kurimo. 2025. Nordparl-tts: Finnish and swedish tts dataset from parliament speech.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, 3.

OpenAI. 2022. Whisper: Robust speech recognition model. `https://github.com/openai/whisper/blob/main/README.md`. Accessed: 2026-01-16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Ronald C Petersen, Barbara Caracciolo, Carol Brayne, Serge Gauthier, Vesna Jelic, and Laura Fratiglioni. 2014. Mild cognitive impairment: a concept in evolution. *J Intern Med.*, 275:214–228.

Bornali Phukon, Xiuwen Zheng, and Mark Hasegawa-Johnson. 2025. Aligning asr evaluation with human and llm judgments: Intelligibility metrics using phonetic, semantic, and nli approaches. In *Proceedings of the 26th Interspeech*, pages 5708–5712, Rotterdam, The Netherlands. Association for Computational Linguistics.

Aurélie Pistono, Jérémie Pariente, and Mélanie Jucla. 2024. Disfluency patterns in alzheimer's disease and frontotemporal lobar degeneration. *Clinical Linguistics & Phonetics*, 38(4):345–358.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Farzana Shahla, Deshpande Ashwin, and Natalie Parde. 2022. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48, Dublin, Ireland. Association for Computational Linguistics.

Leonora Vesterbacka, Faton Rekathati, Robin Kurtz, Justyna Sikora, and Agnes Toftgård. 2025. Swedish whispers; leveraging a massive speech corpus for swedish speech recognition. In *Proceedings of the Interspeech*, pages 758–762, Rotterdam, The Netherlands.

Anders Wallin, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson, Mårten Carlsson, Erik Olsson, Henrik Zetterberg, Kaj Blennow, Johan Svensson, Annika Öhrfelt, Maria Bjerke, Sindre Rolstad, and Carl Eckerström. 2016. The gothenburg mci study: Design and distribution of alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *J Cereb Blood Flow Metab.*, 36(1):114–131.

12. Language Resource References

Språkbanken Text. 2025a. *Mink*. Språkbanken Text. https://spraakbanken.gu.se/mink/.

Språkbanken Text. 2025b. *sbx-swe-mink_analyses*. Språkbanken Text. https://doi.org/10.23695/r5q1-xa67.

Språkbanken Text. 2025c. *sbx-swe-speech2text-transformers-kb_whisper_mp3*. Språkbanken Text. https://doi.org/10.23695/6nry-qr23.

Språkbanken Text. 2025d. *Sparv*. Språkbanken Text. https://spraakbanken.gu.se/sparv/.