# Speech and Language Biomarkers of Neurodegenerative Conditions: Developing Cross-Linguistically Valid Tools for Automatic Analysis

**Iris Nowenstein[1], Marija Stanojevic[2], Gunnar Örnólfsson[3], María Kristín Jónsdóttir[3], Bill Simpson[2], Jennifer Sorinas Nerin[4], Bryndís Bergþórsdóttir[1], Kristín Hannesdóttir[4], Jekaterina Novikova[2], Jelena Curcic[4]**
[1]University of Iceland, Reykjavík, Iceland,
[2]Winterlight Labs (Cambridge Cognition), Toronto, Canada,
[3]Reykjavík University, Reykjavík, Iceland,
[4]Novartis Biomedical Research, Basel, Switzerland and Cambridge MA, USA
{irisen, brb63}@hi.is
{marija.stanojevic, bill.simpson, jekaterina.novikova}@camcog.com
{gunnaro, mariakj}@ru.is
{jelena.curcic, jennifer.sorinas_nerin, kristin.hannesdottir}@novartis.com

**Abstract**

In the last decade, a rapidly growing body of studies has shown promising results for the automatic detection and extraction of speech and language features as biomarkers of neurodegenerative conditions such as Alzheimer's disease. This has sparked great optimism and the development of various digital health tools, but also warnings regarding the predominance of English in the field and calls for linguistically diverse research as well as global, equitable access to novel clinical instruments. To automatically extract clinically relevant features from transcripts in low-resource languages, two approaches are possible: 1) utilizing a limited range of language-specific tools or 2) translating text to English and then extracting the features. We evaluate these approaches for part-of-speech (POS) rates in transcripts of recorded picture descriptions from a cross-sectional study of Icelandic speakers at different stages of Alzheimer's disease and healthy controls. While the translation method merits further exploration, only a subset of the POS categories show a promising correspondence to the direct extraction from the Icelandic transcripts in our results, indicating that the translation method has to be linguistically validated at the individual POS category level.

**Keywords:** machine translation, language-specific tools, Icelandic, part-of-speech (POS), digital health, speech and language biomarkers, neurodegeneration, Alzheimer's disease, Mild Cognitive Impairment, linguistic diversity

## 1. Background and objectives

When digital health tools rely on advances in Natural Language Processing, there is a risk that these tools will only be available for speakers of high-resource languages. This causes linguistic bias and limitations to the access of healthcare solutions which otherwise have the benefit of being noninvasive, fast and low-cost. This type of limitations is present in the context of research on the automatic extraction of speech and language features for the early detection and monitoring of neurodegenerative conditions such as Alzheimer's disease, a field which has rapidly grown in the past decade (e.g. Fraser et al., 2016, Themistocleous et al., 2018, Fraser et al., 2019a, Petti et al., 2020, Balagopalan et al., 2021, Balagopalan and Novikova, 2021, Robin et al., 2021, Cho et al., 2022, and Ehghaghi et al., 2023). The predominance of English in this area of investigation has sparked calls for global equity in the development of automatic speech and language analysis and "timely actions to counter a looming source of inequity in behavioural neurology" (García et al., 2023). This matter is currently of particular relevance, as the UN Decade of Healthy Ageing (2021–2030) and the WHO Global action plan on the public health response to dementia (2017–2025) take place.

A few different routes are available when developing cross-linguistically valid tools for the automatic extraction and analysis of speech and language features in a clinical context. The most direct approach consists in using a mixture of language-specific and language-universal resources to build automated acoustic and lexical/grammatical pipelines, as has been done for English (e.g. Robin et al., 2021, Cho et al., 2022). For example, Cho et al. (2022) report on the analyses of oral picture descriptions from English speakers with amnestic Alzheimer's disease (aAD) or logopenic variant primary progressive aphasia (lvPPA) as well as healthy controls. In their study, the acoustic pipeline is not language-



*Proceedings of the RaPID-5 @LREC-COLING 2024*, pages 26–33

specific and mainly consists of features extracted with a speech activity detector in addition to pitch-tracking. On the other hand, the lexical pipeline makes use of language-specific resources developed specifically for English to extract features such as words' part-of-speech (POS) category, frequency, semantic ambiguity and age of acquisition. The literature in which these lexical/grammatical features are extracted is arguably even more biased towards English-speaking clinical populations than acoustic-centered work in which (possibly) language-universal markers of decline or disease are analyzed. In the context of Scandinavian languages for example, a substantial body of work targeting automatic linguistic feature extraction for the detection of cognitive decline (mostly within the Gothenburg MCI research study, Wallin et al., 2016) has emerged for Swedish, but not other Scandinavian languages to the best of our knowledge. Although some of the earliest work targets acoustic features exclusively (Themistocleous et al., 2018, see also Themistocleous et al., 2020), a number of Swedish MCI studies combined the analysis of acoustic and lexical/grammatical features (e.g. Fraser et al., 2019a, Antonsson et al., 2021) and others focused exclusively on lexical/grammatical features (Fraser et al., 2019b). In all the Swedish MCI studies, the most feasible route of extracting the linguistic features directly from the transcripts was taken, but such an approach depends on the availability of the necessary NLP tools in the language.

Since it is clear that using lexical/grammatical features has the potential to significantly improve disease prediction (Fraser et al., 2019a, Petti et al., 2020, Robin et al., 2021, Cho et al., 2022, Toto et al., 2021), it is imperative to ensure that these features can be extracted from clinical language sample transcripts in under-resourced languages as well. However, the access to the necessary language-specific tools is often limited or non-existent in low-resource languages. This may be remedied by developing and validating specific resources for low-resource languages, but another possible option is the analysis of text samples via an initial automatic translation to English. Such a method has a few obvious advantages and disadvantages. The advantage is that the translation method opens up access to the array of analytical tools developed for English, with some of them showing very promising results for the early detection and monitoring of neurodegenerative diseases (Fraser et al., 2016, Mueller et al., 2018, Petti et al., 2020, Robin et al., 2021, Cho et al., 2022).

The main disadvantages, on the other hand, can be put in two categories. First, the translation of the language samples makes the analysis indirect and therefore more prone to various types of errors and data noise. This includes errors in automatic translation and inaccuracies due to the inevitable non-exact correspondences of the structure of different languages, which might be exacerbated by increased typological distance. This is related to the second disadvantage, which is the partly language-specific nature of disease manifestation. For example, a number of studies have shown an increase in the rate of pronouns and a decrease in the rate of nouns in English speakers with Alzheimer's Disease (Petti et al., 2020, Robin et al., 2021, Cho et al., 2022), but the reverse pattern (decreased pronominal use) has been found in pro-drop languages such as Bengali (Bose et al., 2021) where pronouns are more frequently omitted by Alzheimer's patients. Bengali also has extensive case marking which has largely disappeared from English (McFadden, 2020), and a decreased use of case markers also appears to characterize the language use of Bengali speakers with Alzheimer's (Bose et al., 2021). A translation from Bengali to English would entail the adding of pronouns and loss of case marking, potentially blurring markers of disease. In other words, the translation itself might erase relevant linguistic biomarkers which were present in the original transcript.

Still, the necessity to develop approaches which potentially create more extensive access to linguistic digital health tools as fast as possible amply justifies investigating the potential of the translation method, especially given recent developments in multilingual translation based on foundation models. In light of this, the objective of the present study is to compare POS rates extracted directly and indirectly (through machine translation) from clinical language samples collected from speakers of Icelandic, a low- to medium-resource Germanic language which is related to English but significantly differs from it in various aspects, including a rich case marking system. Icelandic therefore constitutes interesting testing grounds for various reasons, but it is important to note that the vast majority of the world's languages are under-resourced and do not have existing POS taggers or even sufficient data to support machine translation. If the ultimate goal is to develop NLP digital health tools which are globally accessible, the broader endeavor must also include solutions for under-resourced languages. One possible approach to this problem would be concentrating efforts on discovering features which are generalizable across languages (see Lindsay et al., 2021 for such a study with English and French data).

## 2. Methods

To reach our objective, we analyzed oral picture description data from a cross-sectional, noninterventional study conducted at the Memory Clinic of the National University Hospital of Iceland (Curcic et al., 2022) using an Icelandic POS tagger (Jónsson and Loftsson, 2021) and compared the results to Universal Dependency (UD) POS tags (Petrov et al., 2011) extracted from an automatically translated English version of the transcripts.

### 2.1. Participants

Participants in the original study (Curcic et al., 2022) were grouped into four cohorts: 1) cognitively healthy controls (amyloid-negative) without pre-symptomatic biomarkers of Alzheimer's disease, 2) cognitively healthy (amyloid-positive) cohort with pre-symptomatic biomarkers of Alzheimer's disease, 3) people diagnosed with Mild Cognitive Impairment (pre-dementia) and 4) people diagnosed with mild Alzheimer's disease. All participants were aged between 60 and 80 years. The picture description data analyzed in the current study were collected from a total of 48 participants, 12 (25%) were controls, 12 (25%) were pre-symptomatic and 24 (50%) were pre-dementia or had mild Alzheimer's dementia. Although this is the first study in which an Icelandic-specific NLP tool is used to analyze clinical language samples, and the first study in which language features of neurodegeneration are studied in an Icelandic clinical population, we do not analyze participants' POS rates based on their specific cohorts in this particular paper, as the purpose is to evaluate the validity of the machine translation extraction method and compare it to direct feature extraction.

### 2.2. Picture Description Task

The oral picture description data were collected using the Winterlight Speech Assessment, which was developed to record and analyze naturalistic language samples using an app on a tablet. The data set consists of seven different picture descriptions for each individual, recorded in three different sessions if participants completed the protocol: One baseline session with three picture descriptions (conducted in the morning), a follow-up session in the morning four to 32 days later, with two picture descriptions, and an evening session (to produce cognitive fatigue) on the same day as the first follow-up, with two picture descriptions. This creates an unusually robust amount of data per participant, as comparable studies commonly analyze data from a single picture description (e.g. Mueller et al., 2018 and Cho et al., 2022). The seven pictures described are line drawings of scenes specifically conceived to elicit speech for clinical analysis, including the widely used Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass and Wingfield, 1983) as the first stimulus. The participants' speech was recorded through the tablet's microphone and later manually transcribed by a native speaker. The final dataset includes 608 speech samples across 320 picture descriptions from 48 participants, reaching a total of 12 hours and 51 minutes over 73012 word tokens with a mean of two minutes and 25 seconds and 228 word tokens per picture description. No participant is associated with less than three picture descriptions but five descriptions were missing from the dataset and 11 had not been transcribed at the time of analysis.

### 2.3. POS Tagging and Machine Translation

The Icelandic transcripts are POS tagged using ABLTagger, version 3.1 (Jónsson and Loftsson, 2021, Steingrímsson et al., 2019). The tagger is trained on the manually tagged MIM-Gold corpus (Loftsson et al., 2010) and reports a 97.8% cross-validation accuracy on the same corpus, using a fine-grained POS tagset.[1]

To extract features from English, the Icelandic transcripts were translated with the No Language Left Behind (NLLB) model (Costa-jussà et al., 2022). NLLB addresses the translation performance gap between high-resource and low-resource languages by enabling translation across 200 languages and improving translation quality by an average of 44%. The NLLB model was selected because it is open-source and multilingual and therefore fits the premises of the translation method tested in the present paper, but it is important to note that its Icelandic-English translation quality has not been thoroughly evaluated (but see various metrics in Costa-jussà et al. (2022)) and that various other available machine translation tools and large language models, either commercial or not multilingual, should yield higher translation quality (e.g. Google Translate, GPT-4 and Miðeind Vélþýðing[2]). In future work, an important addition to this line of research would be comparing the results across different machine translation tools and evaluating their quality in the context of clinical language samples.

POS tags were extracted from the NLLB translated transcripts with the Spacy library[3] using UD POS tags.[4] The UD POS tagger utilizes a maximum entropy model trained on diverse corpora, demonstrating high accuracy in POS tagging for English. The tagsets used by the Icelandic ABLT-

[1] https://github.com/cadia-lvl/POS
[2] https://huggingface.co/mideind/nmt-doc-en-is-2022-10
[3] https://spacy.io/
[4] https://github.com/explosion/spaCy/blob/master/spacy/glossary.py

agger and the UD framework differ in significant respects. For example, the Icelandic tagset does not derive different tags for auxiliaries which are therefore grouped with verbs in our analysis. Similarly, the infinitival marker *að* 'to' is included in the conjunction category of the Icelandic tagset but was dropped from the category in the present comparison to match the sum of UD POS coordinating conjunction (CCONJ) and subordinating conjunction (SCONJ). Additionally, we do not compare the respective adverb categories which were deemed too incompatible.

The statistical comparison is based on POS category rates for nouns, numerals, verbs, pronouns, conjunctions, prepositions and adjectives. We normalized the rates using the number of intelligible words in the respective transcripts and compared the means of category rates extracted from the Icelandic vs English transcripts (1) across the whole data set (i.e., all four cohorts and all seven pictures) and (2) using paired comparisons for the rates of individual participants across all seven pictures (t-tests and rank correlations). The results are presented in Table 1 and Figure 1 and are further explained and discussed in Sections 3 and 4.

## 3. Results

### 3.1. Mean values across the data set and paired analyses

Table 1 shows the normalized POS category mean values across the whole data set (320 samples from 48 individuals), comparing the tag rates based on extraction (1) directly from the Icelandic transcripts and (2) from a machine translated version of the language samples to English. Table 1 also includes results from paired t-test analyses using the normalized POS rates across all picture descriptions for each individual and the 95% confidence interval for the differences between the two extraction methods at the group level. Finally, we include the (Pearson's) correlations between individuals' ranks in normalized POS rates using the two methods (1-48).

In this analysis, two key results emerge. First, the two different methods (direct and translation) yield very comparable mean rates across the whole data set, with the minimum difference being 0.3% in the case of the numerals and the maximum difference being 3.9% in the case of the conjunctions. Note that the translation method yields consistently lower rates than the direct method. This should be explored further in future work but is possibly in part due to tagset differences and machine translation quality. The second key result is that despite very small differences in mean rates across the whole data set, individuals' rates reveal statistically significant differences for all categories when using a Bonferroni adjusted p-value ($p < 0.007$).

This does not come as a surprise when the individual values across methods are visualized as in Figure 1. The gray lines join together the two data points of each participant, meaning that a preservation of rank across conditions (direct and translation) would result in graphs with no line overlap. As can be seen, the overlap varies greatly between POS categories, reflecting varying amounts of rank differences and a non-systematic lack of equivalence in POS rates. The nouns show the smallest difference in rank correspondence and the greatest discrepancies appear with the adjectives.

To illustrate this further, only 3/48 participants have a rank difference greater than five in the noun category, while this number reaches 25/48 for the adjectives. For example, the speaker with the highest noun rate (rank 1) with the directly extracted features also has the highest rate of nouns with the features extracted from the machine translation method. The correlations in Table 1 reflect this difference in correspondence between POS categories, with the noun category showing the highest correlation between translation and direct features (0.981) while the lowest correlation appears with the adjectives (0.720) and prepositions (0.730). These patterns need to be investigated further in an in-depth analysis of the equivalences between the original transcripts and translations with tagset differences in mind, but they are interesting considering various linguistic factors in the comparison of the Icelandic-English language pair. For example, English and Icelandic share various superficial properties of word order and argument structure, something which should create equivalences in the number of nouns, but the Icelandic case marking system should entail less correspondences in terms of the presence of prepositions. Additionally, a contributing factor might be the size (in tokens) of the different categories, with more robust categories such as nouns (22.5% of the data) being less sensitive to machine translation errors (such as the ones discussed in Subsection 3.2) as compared to adjectives (3.8% of the data).

Given the non-exact nature of translations between languages, it could be furthermore argued that rate differences are less important than rank correspondence for potential clinical markers in a data set of cohorts with varying symptom levels. From this perspective, the feasibility of the translation method varies greatly between POS categories for the Icelandic-English language pair, with the nouns showing the most promising similarities. This is particularly interesting considering evidence from previous research which indicates that noun rate can distinguish between patients with Alzheimer's disease and healthy controls (e.g. Petti

| Category | Direct | Translation | Difference (95%) | t-value | p-value | Rank corr. |
|---|---|---|---|---|---|---|
| **Nouns** | 0.225 | 0.201 | 0.019-0.030 | 8.76 | <0.001 | 0.981 |
| **Numerals** | 0.022 | 0.019 | 0.001-0.003 | 4.24 | <0.001 | 0.834 |
| **Verbs** | 0.186 | 0.169 | 0.014-0.019 | 13.5 | <0.001 | 0.930 |
| **Pronouns** | 0.121 | 0.115 | 0.003-0.011 | 3.62 | <0.001 | 0.871 |
| **Conjunctions** | 0.121 | 0.082 | 0.035-0.041 | 25.94 | <0.001 | 0.921 |
| **Prepositions** | 0.108 | 0.101 | 0.001-0.012 | 2.54 | 0.014 | 0.730 |
| **Adjectives** | 0.038 | 0.031 | 0.005-0.008 | 8.44 | <0.001 | 0.720 |

Table 1: Mean values across the dataset for the direct and translation methods and paired t-test results by individual participant as well as Pearson's correlations for the individual rate ranks (all $p < 0.001$), *N*=$48$.

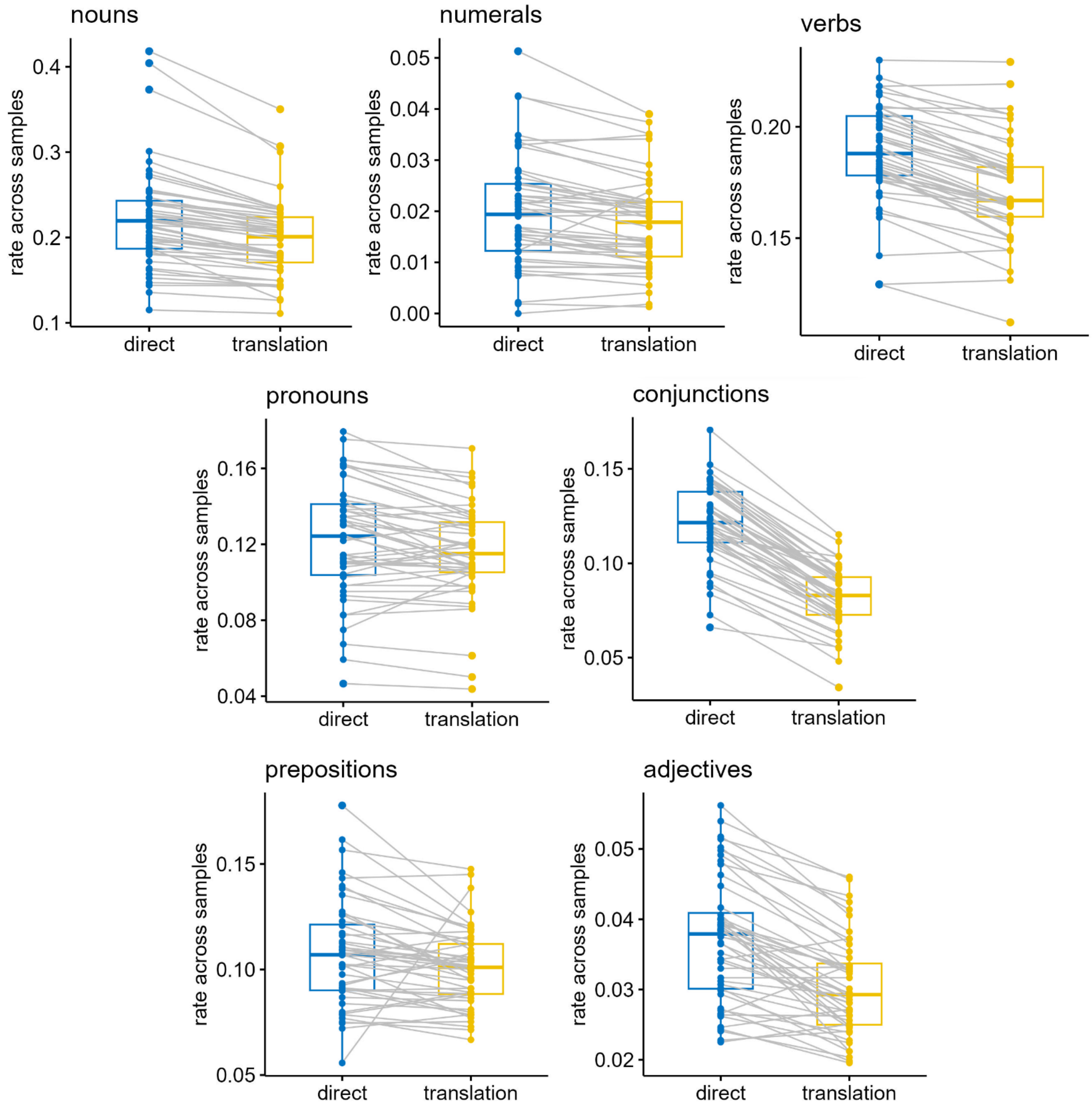


Figure 1: POS rates using the direct and translation methods, *N*=$48$. Distribution of the data and relative position of the individual participants based on their POS category rates.

et al., 2020, Cho et al., 2022). It still is important to stress that although the values of this POS category seem to be well-preserved in machine translation between Icelandic and English, this might not be the case for a language pair with more typological distance. For example, if Mandarin Chinese and English were to be compared, the analysis would have to take into account that objects (including nominal ones) are regularly dropped in Mandarin Chinese (Liu, 2014).

### 3.2. Qualitative observations

To further explore the possible explanations for the numerical discrepancies between methods (direct vs translation) in individuals, we analyzed a small sample of Icelandic transcripts and their translated English versions, focusing on the speakers showing the greatest differences. In this analysis, the bottleneck of machine translation quality became very clear.

For example, one translation included 17 repetitions of the string *and the coffee table* while the original transcript only had a single occurrence of the corresponding *sófaborð* "coffee table". The same translation completely omitted a 16-word passage from the original transcript. Another type of error appeared in the translation of the string *allavegana* "anyway", usually spelled *alla vegana*, which was translated as *all vegans*. These are therefore errors which can both affect POS rates but also the extraction of e.g. word frequency, which additionally differs across languages and cultures.

This brings us to the last observation of machine translation errors, where the original transcript is *eða einhverjir (pause) eitthvað grænmeti* "or some [masculine plural form] (pause) some [correct neuter singular form] vegetables" and the translated version consists of *or some of them might be vegetables*. Here, the machine translation blurs possible disease manifestations such as the repetition, with some of them possibly being language-specific. In this case, the participant initially uses the morphologically inappropriate masculine plural form before correcting themselves and using the neuter singular, in agreement with the word *grænmeti* "vegetable". Indeed, Icelandic has unusually robust nominal concord (Norris, 2012) which could be argued to tax working memory capacity (Hartsuiker and Barkhuysen, 2006). In English, there is only one possible form of the word *some* and therefore no potential for agreement errors. This further illustrates the fact that the development of NLP digital health tools for the diagnosis and monitoring of diseases and disorders based on people's language behavior must take into account possible language-specific manifestations of the conditions being investigated.

## 4. Conclusion

Using a corpus of picture descriptions from Icelandic speakers at different stages of Alzheimer's disease as well as healthy controls, we compared POS feature extraction using (1) the Icelandic transcripts directly and (2) an initial machine translation of the text to English. The results reveal that the use of translated language samples for clinical speech and language analysis has to be linguistically validated at various steps of the process, including the initial automatic translation.

The analysis showed promising similarities between the two methods for a subset of the POS categories, with the most robust individual consistency appearing with nouns. We conclude that the translation method is an avenue which should be further explored, along with the continued development of language-specific tools and detailed work on the manifestations of neurodegenerative diseases across languages. A crucial aspect of deploying computational linguistics methods for the health sector is addressing inequalities in patients' access to cutting-edge NLP digital health tools based on the language they speak. Efforts should be made to address this issue in research.

We leave a clinical cohort classification analysis to future work, as the objective of this paper is an initial linguistically motivated validation of the translation method. Without such a step, it would be impossible to appropriately interpret the success or failure of patient group classification using the two types of feature extraction methods. Additionally, the extraction of various other acoustic and lexical/grammatical features from the dataset is in progress, as well as perceptual clinical ratings by speech-language pathologists. We believe such annotations could contribute to bridging "a growing gulf" (Lindsay et al., 2021) between automatically extracted speech and language features and what is observable by clinicians and people living with neurodegenerative conditions.

## 5. Bibliographical References

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 in Aging Neuroscience*, 12.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in Aging Neuroscience*, 13.

Aparna Balagopalan and Jekaterina Novikova. 2021. Comparing Acoustic-Based Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech 2021*, pages 3800–3804.

Arpita Bose, Niladri S. Dash, Samrah Ahmed, Manaswita Dutta, Aparna Dutt, Ranita Nandi, Yesi

Cheng, and Tina M. D. Mello. 2021. Connected Speech Characteristics of Bengali Speakers With Alzheimer's Disease: Evidence for Language-Specific Diagnostic Markers. *Frontiers in Aging Neuroscience*, 13.

Sunghye Cho, Katheryn Alexandra Quilico Cousins, Sanjana Shellikeri, Sharon Ash, David John Irwin, Mark Yoffe Liberman, Murray Grossman, and Naomi Nevler. 2022. Lexical and Acoustic Speech Features Relating to Alzheimer Disease Pathology. *Neurology*, 99(4):e313–e322.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jelena Curcic, Vanessa Vallejo, Jennifer Sorinas, Oleksandr Sverdlov, Jens Praestgaard, Mateusz Piksa, Mark Deurinck, Gul Erdemli, Maximilian Bügler, Ioannis Tarnanas, Nick Taptiklis, Francesca Cormack, Rebekka Anker, Fabien Massé, William Souillard-Mandar, Nathan Intrator, Lior Molcho, Erica Madero, Nicholas Bott, Mieko Chambers, Josef Tamory, Matias Shulz, Gerardo Fernandez, William Simpson, Jessica Robin, Jón G. Snædal, Jang-Ho Cha, and Kristin Hannesdottir. 2022. Description of the Method for Evaluating Digital Endpoints in Alzheimer Disease Study: Protocol for an Exploratory, Cross-sectional Study. *JMIR research protocols*, 11(8):e35442.

Malikeh Ehghaghi, Marija Stanojevic, Ali Akram, and Jekaterina Novikova. 2023. Factors Affecting the Performance of Automated Speaker Verification in Alzheimer's Disease Clinical Trials. *5th Clinical Natural Language Processing Workshop (ClinicalNLP) at ACL 2023*.

Kathleen C. Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. 2019a. Predicting MCI Status from Multimodal Language Data Using Cascaded Classifiers. *Frontiers in Aging Neuroscience*, 11.

Kathleen C. Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2019b. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech Language*, 53:121–139.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Adolfo M García, Jessica de Leon, Boon Lead Tee, Damián E Blasi, and Maria Luisa Gorno-Tempini. 2023. Speech and language markers of neurodegeneration: a call for global equity. *Brain*, page awad253.

Harold Goodglass and Arthur Wingfield. 1983. *The assessment of aphasia and related disorders*. Lea and Febiger.

Robert J. Hartsuiker and Pashiera N. Barkhuysen. 2006. Language production and working memory: The case of subject-verb agreement. *Language and Cognitive Processes*, 21(1-3):181–204.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021. Language impairment in alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in Aging Neuroscience*, 13.

Chi-Ming Liu. 2014. *A Modular Theory of Radical Pro Drop*. Ph.D. thesis.

Hrafn Loftsson, Jökull H Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a pos-tagged corpus using existing tools. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages, LREC 2010*, page 53.

Thomas McFadden. 2020. *Case in Germanic*, Cambridge Handbooks in Language and Linguistics, page 282–312. Cambridge University Press.

Thomas Melistas, Lefteris Kapelonis, Nikos Antoniou, Petros Mitseas, Dimitris Sgouropoulos, Theodoros Giannakopoulos, Athanasios Katsamanis, and Shrikanth Narayanan. 2023. Cross-Lingual Features for Alzheimer's Dementia Detection from Speech. In *INTERSPEECH 2023*, pages 3008–3012. ISCA.

Kimberly D Mueller, Bruce Hermann, Jonilda Mecollari, and Lyn S Turkstra. 2018. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40(9):917–939.

Mark Norris. 2012. Towards an analysis of concord (in icelandic). In *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pages 205–213. Cascadilla Proceedings.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association: JAMIA*, 27(11):1784–1797.

Jessica Robin, Mengdan Xu, Liam D. Kaufman, and William Simpson. 2021. Using Digital Speech Assessments to Detect Early Signs of Cognitive Impairment. *Frontiers in Digital Health*, 3.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria.

Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. 2020. Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *Plos one*, 15(7):e0236009.

Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. 2018. Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. *Frontiers in Neurology*, 9.

Ermal Toto, ML Tlachac, and Elke A Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4145–4154.

Anders Wallin, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson, et al. 2016. The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism*, 36(1):114–131.

## 6. Language Resource References

Jónsson, Haukur and Loftsson, Hrafn. 2021. *ABLTagger (Lemmatizer) - 3.1.0*. PID http://hdl.handle.net/20.500.12537/134. CLARIN-IS.