The ACoDe Project: Creating a Dementia Corpus for Icelandic

Elena Callegari
University of Iceland
Reykjavík, Iceland
ecallegari@hi.is

Agnes Sólmundsdóttir
University of Iceland
Reykjavík, Iceland
ags46@hi.is

Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

We are creating the very first Dementia corpus for the Icelandic language. Our corpus will contain manually transcribed speech samples elicited from individuals of Icelandic nationality who are aged 60 to 80 and who are suffering from various degrees of Alzheimer's disease. In this paper, we describe our speech elicitation protocol, how we collect the data and how we are transcribing the samples. By sharing our methodology, we hope to spark interest in cross-linguistic research collaborations to develop comparable corpora for languages other than Icelandic.

1 Alzheimer's Disease in Iceland

Alzheimer's disease (henceforth, *AD*) is a type of neurodegenerative disease that causes a progressive decline in cognitive faculties such as memory, decision making and language; around 25 million individuals across the world suffer from AD (Qiu et al., 2022). In Iceland, AD is a particularly pressing concern: according to a 2016 study (Jakobsdottir et al., 2016), people of Icelandic heritage are more likely than other European populations to carry a genetic mutation that results in a greater risk of developing Alzheimer's disease. Moreover, according to data released by the World Health Organization for the year 2019 [1], in Iceland, AD and other dementias were the top cause of death for women and the second cause of death for men. There is currently no cure for AD: there are only therapies that can treat its symptoms and possibly slow its progression. A timely Alzheimers diagnosis provides patients with a better chance of benefiting from existing treatments, with the possibility of accessing support systems and with more time to make plans for the future (Rasmussen and Langerman, 2019); it is, therefore, essential to diagnose this condition as soon as possible. The main procedures currently available to diagnose AD include cognitive tests in combination with PET or MRI, and/or the sampling of cerebrospinal fluid by means of lumbar punctures. These procedures are costly and have long waiting times. This results in delayed diagnoses but also in greater difficulties in monitoring the evolution of the pathology over time.

2 ACoDe: Developing Clinical Speech Analysis for Icelandic

AD affects what we say and how we say it. Multiple studies have shown that individuals suffering from AD exhibit difficulties with word retrieval (Croisile et al., 1996, Kavé and Dassa, 2018), produce fewer information units and content words (Ahmed et al., 2013, Croisile et al., 1996, Kavé and Dassa, 2018), and use more pronouns than healthy age-matched controls (Kavé and Dassa, 2018). Changes to language are already detectable when individuals are diagnosed with Mild Cognitive Impairment (Kavé and Dassa, 2018), a stage of the disease that can occur up to 8 years before the onset of mild Alzheimer's dementia. Spoken language can thus offer a universal and accessible means for measuring neurological health and diagnosing early-stage AD. Indeed, automatic feature extraction and analysis of spoken language for clinical purposes have been attempted before, with remarkable results (Fraser et al., 2014, Peintner et al., 2008, i.a.). Nothing of the sort however currently exists for Icelandic. The ACoDe project ("**A**ssessing **Co**gnitive **De**cline using automatic language analysis") seeks to remedy this gap: our goal

[1] https://data.who.int/countries/352

is to develop software specifically designed for the Icelandic language, aimed at diagnosing Alzheimers disease through automated language analysis. In order to do that, we are collecting speech samples from Icelandic individuals suffering from various stages of AD as well as from healthy, age-matched individuals. Once all participants have been tested, we will train different classifiers on the resulting dataset, to determine whether the classifiers can distinguish between patients and controls, between different stages of AD, and with which accuracy. Our plan is to release the transcriptions in the form of a publicly accessible dataset, so that any researcher working on AD, clinical applications for NLP, or both, may also make use of the data we are collecting. In this paper we describe the speech sample collection process, how we are transcribing the data and how we plan on anonymizing it.

Although our corpus is still in the development phase, we are excited to share our methodology as it is our hope that this will act as a catalyst for future cross-linguistic collaborative research efforts. We are actively seeking partnerships for a pan-European initiative to develop comparable corpora for languages other than Icelandic. We believe that a pan-European project of this type would provide a breakthrough understanding of the effects of Alzheimer's disease on language. For example, analyzing the speech and language patterns of individuals with AD in multiple languages would help confirm the universality or specificity of certain diagnostic markers, possibly leading to breakthroughs in both diagnosis and treatment. Moreover, by working collaboratively on a more global scale, resources (both human and computational) could be shared more efficiently, speeding up the time it takes to compile sufficient data and thus to reach meaningful conclusions.

2.1 Innovative Value

Ours will be the very first dementia dataset for Icelandic. Most of the existing work on clinical feature extraction from language has been done for English (Williams et al., 2021). English also dominates clinical language datasets (MacWhinney et al., 2011), so understanding the extent to which language deterioration due to brain disease generalizes across languages is of great interest. In this respect, Icelandic is an excellent addition to the research effort: Icelandic has a complex inflectional morphology, V2 in embedded clauses (only Yiddish also has embedded V2), as well as other linguistically interesting properties (Thráinsson, 2007). Thus studying how AD affects a language like Icelandic provides invaluable information to better understand how AD affects language more in general.

3 Status & Participants

The ACoDe project officially started in mid 2022. Up until now, we have collected speech samples from more than 50 individuals; we expect to complete the speech collection process by the end of 2024.

We plan on recruiting a total of 120 individuals: 30 patients suffering from Mild Alzheimer's Dementia (MD), 30 patients suffering from Mild Cognitive Decline (MCD) and 30 patients suffering from Subjective Cognitive Decline (SCD). Finally, we will also include 30 healthy controls. There will thus be 3 diagnostic groups and 1 control group, for a total of 4 research groups. Patients with SCD complain of memory problems and overall reduced cognitive abilities, but the extent of these disorders is not such as to be detected by standardized cognitive tests (hence the label *subjective*). Several studies have shown an association between SCD and an increased risk of developing various forms of dementia (Jonker et al., 2000, Geerlings et al., 1999, Jessen et al., 2014, Jessen et al., 2010), hence our interest in this condition. The diagnosis of MCI, MD or SCD is made by qualified clinicians from the Memory Clinic in Reykjavík. All participants will be between the age of 60 and 80. The exclusion criteria (for both patients and controls) are a primary diagnosis of depression of moderate or severe degree, bipolar disorder, schizophrenia, a previous physical brain injury, a neurological disorder or other serious medical condition, a personal history of drug addiction within the past 20 years, issues with alcohol addiction within the past 20 years, the use of antidepressants and the use of benzodiazepine-based sleep medications. To avoid potential confounding factors due to the knowledge of a second language, we are also only accepting individuals who are monolingual speakers of Icelandic. Our study received approval from the Icelandic Research Ethics Committee (*Vísindasiðanefnd*) in September 2021.

We are committed to achieving gender balance across study groups whenever this is possible. In the control group, we have successfully attained an equal gender distribution, with 15 males and 15 females participating in our study. Achieving such balance is more challenging in the patient groups, as the pool of possible participants is much smaller.

3.1 Speech Elicitation Protocol

Each participant is asked to describe in detail: (i) the "picnic scene" by The Arizona Alzheimer's Disease Center. This is a black-and-white depiction of a picnic by the lake; (ii) how they would plan a trip to Akureyri, a city in the north of Iceland; (iii) their childhood home. We decided to include more than the traditional picture-description task, used in many studies on AD, because of evidence that picture-description tasks may not accurately reflect the conversational abilities of individuals with AD (Sajjadi et al., 2012). We chose to include the planning-a-trip kind of narrative task following (Harris et al., 2008), who included a "Plan a trip to New York" question in their own study. According to (Harris et al., 2008), the planning-of-a-trip scenario is complex enough to detect differences between healthy controls and individuals with cognitive decline. This kind of narrative is also effective at engaging episodic memory, which is impaired in individuals suffering from AD (Economou et al., 2016). Finally, we chose to ask participants to describe their childhood home because we reasoned that this question is likely to elicit long responses, minimizing the need for the interviewer to prompt the interviewee with additional follow-up questions to increase the total length of the speech sample. Note also that the description of the childhood home focuses on past events, the trip-planning pertains to future activities, and the description of the picnic scene is not anchored to a specific time. Therefore, this approach could also offer insights into how the disease affects cognitive processing of temporal events and the linguistic expression of time.

The order in which the three main prompts are presented is rotated across participants to mitigate the effect of fatigue on verbal performance. During interviews, participants are encouraged to speak freely and uninterrupted while being audio-recorded. The interviewer uses nonverbal cues and encouraging feedback to make the conversation feel as natural as possible. The goal is to elicit 15 minutes of spoken recording from each participant, and if necessary, pre-decided follow-up questions are asked to elicit longer speech or keep the discussion on topic. The interviewer generally waits 10 seconds after the interviewee has finished speaking before asking follow-up questions. We strive to always ask the follow-up questions in the same order, so as to ensure that speech samples from different participants are maximally comparable. However, we may adjust the question order to maintain a natural conversation, especially if an answer to any of the sub-questions has already been provided earlier.

3.2 Data Anonymization

Participants sign an informed consent at the beginning of the interview which states that all of the participants' identifiable personal information is confidential. Each participant is assigned a research number under which all their data produced in the study is stored. The only link between participant name and research number is kept in an encrypted file which can only be accessed by the interviewer and the researchers at the Memory Clinic who conduct the EEG and cognitive tests. Despite all participants' data being stored anonymously, some identifiable and traceable information can appear during interviews. This is to be expected particularly during the prompt on the participants' childhood home, which often leads to descriptions of family members, and to mentions to schools, specific places and organizations. As Iceland is a fairly low-populated community, this information can in principle reveal the identity of the participant. All information that is deemed to be traceable will therefore be redacted from the transcription dataset before publication, in a way that makes the interviewee unidentifiable while still providing equivalent lexical information needed for language analysis. There are several ways this can be done. One method, recommended by (Aldridge et al., 2010), is to replace potential identifiers such as names, places or organizations with unique identifiers (e.g., pseudonyms), rather than anonymous placeholders (e.g., Person Name). We intend to follow the anonymization guidelines detailed in (Francopoulo and Schaub, 2020). However, owing to Iceland's small population, we must also modify certain phrases that are not specified in the guidelines but could make individuals identifiable in smaller communities. Such elements may include, but are not limited to, names of schools, cities, towns, regions, individuals, and specific dates. We intend to follow the method by which original items are replaced with pseudonyms -or made-up dates in the case of dates-, therefore retaining all grammatical information while protecting anonymity. For example, the fragment sentence in 1, which discusses a local Icelandic school, would be published as 2, where the original school name has been replaced with a pseudonym, keeping the grammatical properties of the original sentence.

(1) ***Já ég var í sa- sama skólanum ee í Langholtsskóla***
Yes I was in sa- same school uh in Langholtsskóli-DAT.
'Yes I went to the same school, Langholtsskóli (A school in Reykjavík).'

(2) ***Já ég var í sa- sama skólanum ee í Borgarskóla***
Yes I was in sa- same school uh in Borgarskóli-DAT.
'Yes I went to the same school, The City School (A fake school that doesn't exist).'

4 Transcription Protocol

Using transcription methods and guidelines from the *Linguistic Data Consortium at the University of Pennsylvania* (hence, LDC), we generate manual text transcriptions from the recorded speech samples (Glenn et al., 2010). The transcriptions are made using a standard text processor, such as Microsoft Word or TextPad, and exported to plain text files (.txt). The transcriptions contain speech from both speakers, i.e. interviewer and interviewee, and accurately annotate any interjections or overlaps, providing detailed transcriptions of the conversations as a whole. The transcriptions are verbatim and orthographic using standard Icelandic spelling. Filled pauses, false starts, repeated words, repairs, restarts, partial words, spoonerisms, speech errors and speaker noises are all marked and annotated in accordance with the transcription protocol. We follow the LDC guidelines as much as possible with some modifications for Icelandic. These adjustments primarily involve Icelandic discourse particles, which differ from those in English. For instance, we created a list of Icelandic-specific particles, including "uu", "ömm", "sko", and "hérna". The word "hérna" is noteworthy, as it serves as both a lexical word, an adverb of place meaning "here", and a planning marker used to indicate hesitation or to maintain a speaker's turn in a conversation, similar to the English particle "uhm" (Hilmisdóttir, 2011). We therefore annotate this word to differentiate between the two meanings as shown in 3 and 4.

(3) Adverb of Place

Hérna situr par á teppi.
Here sits couple on blanket.
'Here is a couple sitting on a blanket.'

(4) Planning Marker

Og *hérna* það var bara *hérna* mjög gaman.
And *uhm* that was just *uhm* very fun.
'And that was just very fun.'

5 Format

All transcriptions of the speech samples resulting from our project will be made publicly available in the form of a CC-BY 4.0-license corpus, which will be distributed in TEI-conformant format and will be accessible to everyone on the Icelandic CLARIN repository. The corpus will only include the transcriptions, not the audio files, as it is much harder to preserve anonymity with audio files. The released version will include annotation that builds on several other Icelandic CLARIN resources, released via projects that have been carried out in recent years, most notably the Icelandic Language Technology Programme Nikulásdóttir et al., 2020, and build on the experience and protocols accumulated within these projects, ensuring interoperability between systems and readiness of the human resources available in Iceland for future work. Our open-source policy and standardized packaging of our data will encourage the use of our output in future R&D projects across academia and industry.

References

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven alzheimers disease. *Brain*, *136*(12), 3727–3737.

Aldridge, J., Medina, J., & Ralphs, R. (2010). The problem of proliferation: Guidelines for improving the security of qualitative data in a digital age. *Research Ethics*, *6*(1), 3–9.

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, *53*(1), 1–19.

Economou, A., Routsis, C., & Papageorgiou, S. G. (2016). Episodic memory in alzheimer disease, frontotemporal dementia, and dementia with lewy bodies/parkinson disease dementia. *Alzheimer Disease & Associated Disorders*, *30*(1), 47–52.

Francopoulo, G., & Schaub, L.-P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. *workshop on Legal and Ethical Issues (Legal2020)*, 9–14. https://hal.science/hal-02939437

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, *55*, 43–60.

Geerlings, M. I., Jonker, C., Bouter, L. M., Adèr, H. J., & Schmand, B. (1999). Association between memory complaints and incident alzheimers disease in elderly people with normal baseline cognition. *American Journal of Psychiatry*, *156*(4), 531–537.

Glenn, M. L., Strassel, S. M., Lee, H., Maeda, K., Zakhary, R., & Li, X. (2010). Transcription methods for consistency, volume and efficiency. *LREC*.

Harris, J. L., Kiran, S., Marquardt, T. P., & Fleming, V. B. (2008). Communication wellness check-upl': Age-related changes in communicative abilities. *Aphasiology*, *22*(7-8), 813–825.

Hilmisdóttir, H. (2011). Giving a tone of determination: The interactional functions of nú as a tone particle in icelandic conversation. *Journal of Pragmatics*, *43*(1), 261–287.

Jakobsdottir, J., van der Lee, S. J., Bis, J. C., Chouraki, V., Li-Kroeger, D., Yamamoto, S., Grove, M. L., Naj, A., Vronskaya, M., Salazar, J. L., et al. (2016). Rare functional variant in tm2d3 is associated with late-onset alzheimer's disease. *PLoS genetics*, *12*(10), e1006327.

Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K. A., Van Der Flier, W. M., et al. (2014). A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease. *Alzheimer's & dementia*, *10*(6), 844–852.

Jessen, F., Wiese, B., Bachmann, C., Eifflaender-Gorfer, S., Haller, F., Kölsch, H., Luck, T., Mösch, E., van den Bussche, H., Wagner, M., et al. (2010). Prediction of dementia by subjective memory impairment: Effects of severity and temporal association with cognitive impairment. *Archives of general psychiatry*, *67*(4), 414–422.

Jonker, C., Geerlings, M. I., & Schmand, B. (2000). Are memory complaints predictive for dementia? a review of clinical and population-based studies. *International journal of geriatric psychiatry*, *15*(11), 983–991.

Kavé, G., & Dassa, A. (2018). Severity of alzheimers disease and language features in picture descriptions. *Aphasiology*, *32*(1), 27–40.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286–1307.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Steingrmsson, S. (2020). Language technology programme for icelandic 2019-2023. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3414–3422.

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M. L. G., & Ogar, J. (2008). Learning diagnostic models using speech and language measures. *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4648–4651.

Qiu, C., Kivipelto, M., & Von Strauss, E. (2022). Epidemiology of alzheimer's disease: Occurrence, determinants, and strategies toward intervention. *Dialogues in clinical neuroscience*.

Rasmussen, J., & Langerman, H. (2019). Alzheimers disease–why we need early diagnosis. *Degenerative neurological and neuromuscular disease*, 123–130.

Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, *26*(6), 847–866.

Thráinsson, H. (2007). *The syntax of Icelandic*. Cambridge University Press.

Williams, E., McAuliffe, M., & Theys, C. (2021). Language changes in alzheimers disease: A systematic review of verb processing. *Brain and Language*, *223*, 105041.