# Towards Automatic Transcription of ILSE – an Interdisciplinary Longitudinal Study of Adult Development and Aging

**Jochen Weiner[1], Claudia Frankenberg[2], Dominic Telaar[1], Britta Wendelstein[2,3], Johannes Schröder[2,3], Tanja Schultz[1]**

[1]Cognitive Systems Lab, University of Bremen (formerly Karlsruhe Institute of Technology), Germany
[2]Section of Geriatric Psychiatry, University Hospital Heidelberg, Germany
[3]Institute of Gerontology, Universität Heidelberg, Germany
jochen.weiner@uni-bremen.de

**Abstract**

The *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) was created to facilitate the study of challenges posed by rapidly aging societies in developed countries such as Germany. ILSE contains over 8,000 hours of biographic interviews recorded from more than 1,000 participants over the course of 20 years. Investigations on various aspects of aging, such as cognitive decline, often rely on the analysis of linguistic features which can be derived from spoken content like these interviews. However, transcribing speech is a time and cost consuming manual process and so far only 380 hours of ILSE interviews have been transcribed. Thus, it is the aim of our work to establish technical systems to fully automatically transcribe the ILSE interview data. The joint occurrence of poor recording quality, long audio segments, erroneous transcriptions, varying speaking styles & crosstalk, and emotional & dialectal speech in these interviews presents challenges for automatic speech recognition (ASR). We describe our ongoing work towards the fully automatic transcription of all ILSE interviews and the steps we implemented in preparing the transcriptions to meet the interviews' challenges. Using a recursive long audio alignment procedure 96 hours of the transcribed data have been made accessible for ASR training.

**Keywords:** ILSE, ASR corpus, long audio alignment, dementia

## 1. Introduction

The population in developed countries is aging rapidly. In Germany, for example, the most populous age group in the year 1950 were ten-year-olds and in 2000 forty-year-olds; for 2050 it is estimated to be the sixty-year-olds (Statistische Ämter des Bundes und der Länder, 2011). This demographic change is caused by and causes transitions in family structures, way of life, and work force and thus bears tremendous challenges for the society. To tackle these challenges and mitigate the consequences, researchers of various disciplines such as physicians, psychologists, gerontologists, sociologists, linguists, and engineers are joining forces in various ongoing interdisciplinary projects and initiatives. One of these projects is the *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) (Martin and Martin, 2000; Lehr et al., 2000; Schönknecht et al., 2005), a cohort-based longitudinal study similar to the Health and Retirement Study (Juster and Suzman, 1995) and the Victoria Longitudinal Study (Dixon and de Frias, 2004) in the USA, and PAQUID (Dartigues et al., 1992) in France.

Within ILSE, over a thousand subjects participated in up to four biographic interviews over a time span of more than 20 years, resulting in over 8,000 hours of recorded speech. These interviews are complemented by data from psychological, cognitive, physical and dental examinations. Among other indicators, it includes for each participant a diagnosis about the existence and level of cognitive decline. In order to use linguistic features to detect dementia in the participants we require transcripts of these interviews (cf. Wendelstein (2016)). Furthermore, the data contained in ILSE enables future studies in disciplines such as psychology, sociology and linguistics which also require orthographic transcripts with a word by word transcription (Selting et al., 2009; Edwards and Lampert, 1993) and systematic standards for nonverbal information (Wendelstein and Sattler, 2011) to be available. On average, the manual creation of such a basic transcription by a human expert turned out to take more than 12 times real-time, i.e. one human expert would spend several decades transcribing all recordings. Consequently, only a small fraction of the data has been transcribed so far. We intend to leverage methods of automatic speech recognition (ASR) to transcribe the full set of interviews with minimal human effort. Large-scale international programs (e.g. the DARPA EARS and GALE programs on broadcast data), as well as studies like MALACH (Byrne et al., 2004), the National Gallery of the Spoken Word (Hansen et al., 2001), and the Czech Radio archive (Nouza et al., 2014), to name only a few, have shown great progress in fully automatic transcription.

In this paper, we present our ongoing work towards a fully automatic transcription of the ILSE interviews. For this purpose, we first introduce the ILSE study (Section 2) with a focus on the cognitive diagnoses (Section 3) and interview data (Section 4) and then describe the various challenges (Section 5) and proposed solutions (Section 6) to transform the text and audio data from the interviews into a database suitable for speech processing.

## 2. ILSE

For the ILSE study, an interdisciplinary team at Heidelberg University was commissioned to collect a large-scale corpus with the goal to assess healthy and satisfying aging in middle adulthood and later life (Martin et al., 2000). Within this goal, the study was supposed to compare adults from the western and eastern parts of Germany. Over a thousand

participants from the Leipzig region (Eastern Germany) and the Heidelberg/Mannheim region (Western Germany) from two birth cohorts, born 1930-1932 (C30) and 1950-1952 (C50), were randomly selected and recruited from the German community registers. Since registration in these registers is mandatory for German citizens older than 15 years, the recruitment procedure and the number of participants in this study (Table 1) meet the criteria of representativeness for the sampled population (Martin and Martin, 2000).

| | Western Germany | | Eastern Germany | | Total |
|---|---|---|---|---|---|
| | Male | Female | Male | Female | |
| **C30** | 130 | 120 | 130 | 120 | 500 |
| **C50** | 130 | 120 | 130 | 122 | 502 |
| **Total** | 260 | 240 | 260 | 242 | 1002 |

Table 1: The number of ILSE participants by region, gender, and age group

In order to assess the participants' personality, cognitive functioning, subjective well-being, and health ILSE comprises results from psychological, cognitive, physical, and dental examinations, as well as biographic information derived from questionnaires and semi-standardized interviews.

Data collection for ILSE began in 1993 and is still ongoing over 20 years later. To date, four measurements have been conducted in time periods T1 - T4, as summarized in Table 2. Participants were invited to take part in all four measurements. Almost 90% of the participants of the first measurement returned for the second measurement and 65% returned again for the third measurement. Reasons why participants did not return include that they had developed physical or cognitive handicaps that would make participation too troublesome, had died, lost interest, or had moved to other regions of the country.

Given the representative participant group and up to four measurements per participant over a time-span of 20 years, ILSE fosters representative investigations along two dimensions: Firstly, it allows for the detailed longitudinal analysis of individual aging. Secondly, it enables the study of inter-individual differences along the dimensions of region, gender, and age cohort.

| **T** | **T1** | **T2** | **T3** | **T4** |
|---|---|---|---|---|
| **period** | 1993–1996 | 1997–2000 | 2005–2008 | 2013–2016 |
| **age C30** | 61–66 | 65–70 | 73–77 | 81–86 |
| **age C50** | 41–46 | 45–50 | 53–57 | 61–66 |

Table 2: The four ILSE measurements: Time period and age of participants in both age groups (in years).

## 3. ILSE Cognitive Diagnoses

Cognitive diagnoses were established in a two-step consensus process by trained psychiatrists using information from neuropsychological, anamnestic, clinical, and laboratory tests. Neuropsychological functioning was evaluated using a number of well-established tests, including the Mini Mental State Examination (Folstein et al., 1975), the trail making test (Army Individual Test Battery, 1944), the Aufmerksamkeits-Belastungs-Test "d2" (Brickenkamp, 1994), the CERAD's 15-item version of the Boston-Naming-Test (Morris et al., 1989), subtests from the Nuremberg Gerontopsychological Inventory (Oswald and Fleischmann, 1993), the Leistungsprüfsystem (Horn, 1983), the Hamburg-Wechsler Intelligence Test (Tewes, 1991), the Wechsler Memory Scale (Härting et al., 2000), and an adapted version of the Bielefelder Autobiographical Memory Inventory (Fast et al., 2006; Fast et al., 2007).

*Aging-associated cognitive decline (AACD)* was diagnosed if both subjective and objective impairment was found according to criteria by Levy (1994). Subjective impairment required a report by the participant that cognitive function had declined. Objective impairment was identified in one of five cognitive domains (memory/learning, attention/concentration, abstract thinking, language and visuospatial functioning) if participants scored at least one standard deviation below age- and education-adjusted normative levels (cf. Schönknecht et al. (2005)). *Mild cognitive disorder (MCD)* was diagnosed using ICD-10 criteria which require the presence of a cerebral or systemic disorder causing cerebral dysfunction (cf. Schönknecht et al. (2005)). The cognitive impairments in AACD and MCD are less severe than those found in the following two types of dementia: *Alzheimer's disease (AD)* was diagnosed if the NINCDS-ADRDA criteria (McKhann et al., 1984) were met which include impairments in memory, language, perceptual skills, constructive abilities, orientation, problem solving and functional abilities. AD is the most common form of dementia (World Health Organization and Alzheimer's Disease International, 2012), leading to severe impairments of cognitive domains that result in reduced everyday functioning. *Vascular dementia (VAD)*, another common form of dementia, caused by disturbed blood flow in the brain. VAD was diagnosed using NINDS-AIREN criteria (Román et al., 1993) which require disorders in amnestic domains as well as in two of the following cognitive domains: orientation, attention, language and expression, visuospatial functions, calculations, executive functions, motor skills, abstract thinking skills. Cognitively healthy participants were classified as *control subjects (CONTR)*.

The resulting diagnoses of the ILSE participants are shown in Table 3. No diagnosis was established for the C50 participants in the first two measurements since these participants were in their forties and therefore extremely unlikely to have developed any of the above cognitive impairments. Being 20 years older, a few C30 participants had already developed AACD or MCD when the first two measurements were taken. In the third measurement the first C30 participants were diagnosed with dementia (AD and VAD). The percentage of dementia diagnoses in the third measurement is about the percentage that is to be expected given the prevalence of dementia in Germany estimated by Alzheimer's Disease international (Prince et al., 2015, p. 20). The diagnoses for the fourth measurement have not yet been finalized since that measurement has not yet been completed.

| Cohort | Diagnosis | T1 | T2 | T3 |
|---|---|---|---|---|
| C30 | AACD | 13.0 % | 23.6 % | 27.7 % |
| | MCD | 5.6 % | 7.8 % | 5.4 % |
| | AD | - | - | 5.4 % |
| | VAD | - | - | 0.6 % |
| | CONTR | 81.4 % | 68.6 % | 60.8 % |
| C50 | AACD | | | 5.6 % |
| | MCD | | | 3.6 % |
| | AD | | | - |
| | VAD | | | - |
| | CONTR | | | 90.8 % |

Table 3: Percentage of each diagnosis in the first three measurements in the two cohorts. No diagnosis was established for the C50 participants in the first two measurements.

## 4. ILSE Interviews

In the biographic interviews each participant was interviewed by one of 53 interviewers. The interviewers followed a semi-standardized interviewing procedure to ensure topic consistency within one measurement: They prompted the participants with prepared short questions and encouraged them to elaborate on details by back-channeling and further queries. With short interviewer questions and detailed participant replies, the interviews are dominated by the participants' speech. Since the participants were given ample time to think about their answer, the interviews also contain large portions of silence. The interviews lasted $6.0 \pm 2.6$ hours in the first measurement, $2.5 \pm 0.7$ hours in the second, $1.8 \pm 0.8$ hours in the third and $1.3 \pm 0.3$ hours in the fourth measurement. The length of the interviews is decreasing over time since fewer questions were asked at later measurements.

All interviews were recorded using a stereo voice recorder sitting on the table between the interviewer and the participant. For the first two measurements analog recording devices were used, which stored the interviews on tape. The tapes were kept in metal lockers until digitalization began in 2008. Since then the majority of these tapes has been digitalized using a sampling rate of 48 kHz, quantization of 16 bit and an uncompressed PCM format. For the third measurement and the first half of the fourth measurement a digital recording device was used which sampled with 44.1 kHz and stored all recordings in MP3 format with bit rates varying between 56 and 160 kb/s, while for the currently ongoing second half of the fourth measurement digital recording devices store all recordings using a sampling rate of 16 kHz, quantization of 16 bit and an uncompressed PCM format.

A total of approximately 8,000 hours of interviews has been recorded to date and less than 5 % have been manually transcribed. These transcripts have been created on a per-participant-basis for a total of 74 participants interviewed by 19 interviewers as shown in Table 4.

| | Western Germany | | Eastern Germany | | Total |
|---|---|---|---|---|---|
| | Male | Female | Male | Female | |
| C30 | 108h (18) | 47h (10) | 102h (19) | 55h (13) | 312h (60) |
| C50 | 32h (6) | 19h (3) | 7h (1) | 14h (4) | 72h (14) |
| Total | 140h (24) | 66h (13) | 109h (20) | 69h (17) | 384h (74) |

Table 4: Amount of transcribed data in hours and (in parentheses) the number of participants whose interviews were manually transcribed.

## 5. Challenges

The ILSE interviews and their transcriptions pose several challenges for automatic speech and language processing. While most of these issues have already been addressed in research and development, their joint occurrence makes the task of automatic speech recognition (ASR) for ILSE quite demanding:

**Transcription Quality:** The transcriptions reflect the spoken content of a complete interview at word surface level. Speaker changes and utterance boundaries were transcribed but not marked in the interview recordings, and no time alignment between recording and transcript was created. Thus, one transcription is provided either per part of the interview (one side of a tape, $\sim$45 min) in T1 and T2 or per full interview (0.3-2 hours) in T3 and T4. A third of the transcripts (35 %, 145 hours) has been created very carefully using a systematic transcription approach and manual quality cross-checks, resulting in an orthographic basic transcription (Wendelstein, 2016). However, the majority of the transcripts (65 %, 236 hours) does not follow the criteria of linguistic transcripts, was created without any post-checks and includes hardly any transcribed vocables (hesitations, back-channeling, disfluencies). Overall, the transcription quality ranges from very reliable to hardly usable.

**Anonymized Transcriptions:** To safeguard the participants' privacy and comply with ethical, legal, and social responsibilities, all transcriptions were anonymized. Proper names, city names, and dates were substituted with generic place holders: e.g. surnames were replaced by the most frequent German surname "Müller". Thus, there is some mismatch between transcript and spoken content.

**Recording Quality:** The original interview setup was created to comfort the participants. Little attention was paid to the audio quality, and automatic speech processing was not envisioned. Therefore, little care was taken concerning e.g. the position and orientation of the microphone placed on the table. As a consequence, the data is rather noisy, including reverberation, table tapping, shuffling paper, writing, placing objects, in addition to environmental noise from the tape recorder, the street, and alike. Furthermore, when digitalization started in 2008 the age of the recording tapes had already reduced the quality of the taped material. Often, speech is hardly intelligible, which is also noted in the transcriptions. Overall, the signal-to-noise ratio (SNR) ranges between -9 dB and 85 dB according to the WADA-SNR algorithm (Kim and Stern, 2008).

**Speaking Style and Crosstalk:** The speaking styles of interviewers and participants differ substantially: While the semi-standardized questions of the interviewers are usually short and well planned, participants answer in detail, great length, and in a very spontaneous fashion. In addition,

crosstalk, disfluencies, and back-channeling occur very frequently.

**Emotional Speech:** The content of the biographic interviews covers a wide collection of topics, ranging from family and career to very private questions like sexual experiences. As a result, major parts of the interviews contain emotional speech with strongly varying arousal and valence (Schlosberg, 1954; Banse and Scherer, 1996).

**Dialectal Speech:** The ILSE interviews took place in Leipzig and Heidelberg/Mannheim, both with very distinct regional dialects. While most parts of the transcriptions are written in the original standardized High-German writing system, other parts were produced using quasi-phonetic transcripts to write dialectal variants. The quasi-phonetic dialect variants were not marked and the standardized form of the dialectal variant was not noted.

## 6. Data Preparation and Baseline

In this section we describe the steps taken to prepare the ILSE interviews for fully automatic speech and language processing. These steps will have to consider all of the challenges described above.

### 6.1. Data Division

In a first step, we divided the transcribed interviews into training, development and test set, as shown in Table 5. Using these sets we will develop and evaluate an ASR system which we will then use to transcribe the full set of (currently untranscribed) ILSE interviews. We divided the data considering the following constraints:

- Disjoint participant groups: no participant's interviews appear in more than one set, while interviewers may appear in more than one set
- Equal gender/age/region distribution: the proportions of genders, cohorts and regions are roughly the same across all sets
- Representativeness and reliability of results: reasonable amounts of participant's speech were places in each set to ensure reliable results, and each set has a reasonable number of different participants to ensure representative results.

| | Training | Development | Test |
|---|---|---|---|
| **participants** | 50 | 10 | 14 |
| **audio length** | 265:23 | 35:52 | 62:00 |

Table 5: The numbers of participants and audio durations [hrs:min] in the training, development and test sets.

### 6.2. Text Processing

The transcripts were created in a non-standardized manner (see Section 5.) resulting in inconsistent file formats and structure. Therefore, the transcriptions were first converted into a consistently structured text format. Various types of annotations such as pauses or hesitations that occur in a few transcripts were analyzed and normalized along with titles, numbers, dates and abbreviations. Anonymizations were identified and labeled accordingly.

#### 6.2.1. Vocabulary

The vocabulary of the normalized transcriptions follows Zipf's law (Zipf, 1949) with over half of the training words occurring just once. Table 6 shows statistics of the vocabulary (word types) and running words (word tokens) in the training, development and test sets. As expected, the out-of-vocabulary (OOV) rate of the training word types for the interviewers' speech is considerably lower than for the participants' speech. This reflects the fact that the interviewers' speech is more planned and that the interview questions largely overlap in style and topic over the course of one measurement.

| | Training | Development | Test |
|---|---|---|---|
| **word types** | 50 k | 15 k | 20 k |
| **word tokens** | 2,000 k | 270 k | 500 k |
| **- interviewers** | 450 k | 70 k | 120 k |
| **- participants** | 1,550 k | 200 k | 380 k |
| **OOV rate** | | 1.82 % | 1.50 % |
| **- interviewers** | | 0.77 % | 0.58 % |
| **- participants** | | 2.18 % | 1.79 % |

Table 6: The number of word types and word tokens in the training, development and test sets and the OOV of the training set vocabulary on the development and test sets.

#### 6.2.2. Language Modeling

We trained statistical baseline language models (LMs) based on the normalized transcriptions from the training set. A 3-gram language model with Kneser-Ney smoothing was trained using the SRILM toolkit (Stolcke, 2002) and a recurrent neural network LM with 200 hidden nodes was trained using the RNNLM toolkit (Mikolov et al., 2011). Perplexities of these language models on the development and test sets are shown in Table 7.

| | Development | Test |
|---|---|---|
| **3-gram LM** | 144.7 | 131.7 |
| **RNN LM** | 130.9 | 118.7 |

Table 7: Perplexity of the language models trained on the training transcriptions (50 k word types and 2,000 k word tokens).

### 6.3. Speech Data Processing

For acoustic model training the audio recordings need to be split into segments of appropriate length with a reasonably good alignment to the corresponding word-level transcriptions. The ILSE corpus originally provides unaligned transcriptions for audio recordings with durations of at least 45 minutes. Neither a flat start on these audio recordings nor aligning them to the transcription with a deep neural network (DNN) model trained with read speech from the German edition (Schultz, 2014) of the GlobalPhone corpus (Schultz et al., 2013) was successful. Reasons for this are the recording length, lack of transcription alignment,

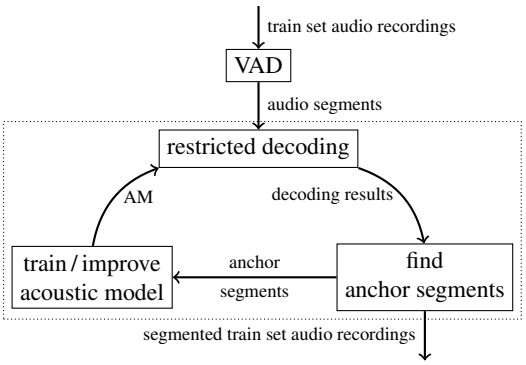


Figure 1: Long audio alignment procedure iteratively increasing the amount of segmented training data.

erroneous transcriptions, and the overall challenging and noisy data with a varying SNR. We therefore implemented a *long audio alignment* procedure (Figure 1) based on the approaches described by Moreno et al. (1998) and Hazen (2006) to segment the training data.

#### 6.3.1. Voice Activity Detection (VAD)

The procedure begins with voice activity detection (VAD). Our VAD system consists of a Hidden-Markov-Model (HMM) recognizer with one model for silence and one model for non-silence. Both were modeled using Gaussian Mixture Models (GMM) with 128 Gaussians each and were trained on the GlobalPhone corpus (Schultz, 2014) using our in-house BioKIT toolkit (Telaar et al., 2014). As features we use Mel-frequency cepstral coefficients (MFCCs) with first and second order derivatives plus zero crossing rate.

We run one decoding pass on the training set with these models, then adapt the models for each audio recording using Maximum Likelihood Linear Regression (MLLR). In a second decoding pass the transformed models are used to label silent and non-silent segments. Finally, the audio recordings are partitioned by splitting all long silent regions ($> 2$ s) in the middle. This procedure ensures that audio recordings are only split at long pauses and that no data is discarded for the next processing step.

#### 6.3.2. Long Audio Alignment

While the VAD procedure splits the audio recording into a number of partitions, the audio recording's transcription remains unpartitioned. Therefore we start a *long audio alignment* procedure which iteratively aligns a larger portion of the audio partitions and thus increases the amount of data available for acoustic model (AM) training.

In each iteration all the partitions produced by the VAD undergo restricted decoding using the BioKIT toolkit: The search vocabulary is restricted to the vocabulary of the audio recording transcription plus fillers. Thus, we can make use of the transcriptions which originally could not successfully be aligned with the audio. After decoding, the recognition hypotheses of all partitions of an audio recording are concatenated and aligned with the transcription reference using Levenshtein distance. Words that have been recognized correctly are considered to be candidates for "anchor segments", i.e. segments in which the transcription can be correctly aligned with the decoding hypothesis. Sequences of candidates which are longer than two words are assumed to be a good transcription of the corresponding audio. These sequences are then split at speaker turns and used to train a new AM for the next iteration. In each iteration, an increasing amount of speech is recognized correctly. Hence, more anchor segment data can be extracted as training material for the next AM which, in turn, will hopefully lead to an AM recognizing a larger amount of speech correctly.

The AMs in this procedure use (stacked) MFCC features and were trained using the Kaldi toolkit (Povey et al., 2011). For the initial AM 2 hours of recordings were manually segmented into segments shorter than two minutes. These segments were labeled using a DNN model trained with read speech from the GlobalPhone corpus (Schultz, 2014) and then a context-dependent GMM triphone model was trained with these labels.

**First iteration:** We used the initial AM and for each audio recording restricted the search vocabulary to the vocabulary of the audio recording transcription plus fillers. Additionally the LM was restricted to a 3-gram LM with Witten-Bell discounting built on the audio recording transcription. From the decoding results we extracted 44 h of anchor segments. We labeled these anchor segments using the initial AM. With these labels we trained a DNN model which was initialized with deep belief network (DBN) pre-training followed by cross-entropy training and state-level minimum Bayes risk (sMBR) sequence-training (Veselý et al., 2013). The network takes a 440-dimensional feature vector of 11 stacked 40-dimensional LDA-transformed stacked MFCCs as input and has 6 hidden layers with 2,048 nodes each and 4,622 nodes in the output layer.

**Second iteration:** The DNN form the first iteration was then used in the second iteration in which we used a 3-gram LM built on the whole training transcriptions but still restricted the search vocabulary to the vocabulary of the audio recording transcription. We extracted 96 h of anchor segments, which corresponds to roughly one third of the whole training data. The average length of these segments is 2.0 s and the average number of words is 5.4 words per segment. In these 96 h of data all interviewers and participants from the training set are represented, where 22 h are interviewer segments and 74 h are participant segments. Decoding the unsegmented development set using an AM trained on these segments, the 3-gram LM and the training vocabulary results in a word error rate (WER) of 67.5%. In the future, this system will serve as a baseline for the development of an ASR system for the fully automatic transcription of the ILSE corpus.

## 7. Discussion and Future Work

By selecting only sequences of correctly recognized words as anchor points we make sure that only segments which have been transcribed correctly are used for AM training. The drawback of this approach is that due to the performance of the baseline system it will not find every segment for which the manual transcription is correct. By training an AM only on the anchor segments, i.e. words that the system has already recognized correctly, the system is reinforced in what it can already recognize and does not learn

anything about the words that are not included in the anchor segments. Therefore, for future work we plan to extend the long audio alignment procedure so that we can also reliably select segments between anchor segments.
In order to have a number of reliable short segments for training and evaluation, a part of the transcribed data is currently also being segmented manually. However, as the manual segmentation takes about 5 times real-time, it will not be able to replace the long audio alignment based segmentation.
The manual segments will, however, enable us to interpret the ASR system's performance in comparison with human performance. Since speech is often hardly intelligible (see Section 5.), it may very well be true that the system's overall performance is not too far away from human transcription performance.
Using the baseline reported in this paper we will approach more of the corpus' challenges, for example training different acoustic models for the two different regional dialects, and denoising the audio recordings.

## 8. Conclusion

We have presented the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE), a corpus of over 8,000 hours of biographic interviews recorded from over a thousand participants in four measurements over the course of 20 years. It is our aim to fully automatically transcribe this corpus using automatic speech recognition and extract linguistic features for the detection of cognitive decline from the resulting transcripts. The joint occurrence of poor recording quality, long audio segments, erroneous transcriptions, varying speaking styles & crosstalk and emotional & dialectal speech in these interviews presents challenges for ASR.
We have described how we prepared the corpus data to take up these challenges. In particular, we have presented corpus analyses and speech data preparation steps. We use a recursive long audio alignment procedure to split the long transcriptions into smaller segments which we can use for acoustic model training. With this procedure we have so far been able to make roughly a third of the data available for training. With this data we have created a baseline ASR system with a WER of 67.5% on the unsegmented development set.
Using this baseline we will work further towards the creation of an ASR system which can fully automatically transcribe the remaining large portion of untranscribed data.

## 9. Acknowledgements

The Interdisciplinary Longitudinal Study of Adult Development and Aging (ILSE) was supported by the Research Program of the State of Baden-Württemberg, the Federal Ministry of Family, Senior Citizens, Women and Youth (AZ: 301-1720-295/2), and the Dietmar-Hopp-Stiftung.

## 10. Bibliographical References

Army Individual Test Battery. (1944). *Manual of Directions and Scoring. War Department, Adjutant General's Office, Washington, DC*.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614–636.

Brickenkamp, R. (1994). *Der Aufmerksamkeits-Belastungs-Test (d2-Test). Handanweisung*. Hogrefe, Göttingen, 8. edition.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., and Zhu, W.-J. (2004). Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435.

Dartigues, J.-F., Gagnon, M., Barberger-Gateau, P., Letenneur, L., Commenges, D., Sauvel, C., Michel, P., and Salamon, R. (1992). The Paquid epidemiological program on brain ageing. *Neuroepidemiology*, 11(Suppl. 1):14–18.

Dixon, R. A. and de Frias, C. M. (2004). The Victoria Longitudinal Study: From Characterizing Cognitive Aging to Illustrating Changes in Memory Compensation. *Aging, Neuropsychology, and Cognition*, 11(2-3):346–376.

Jane A Edwards et al., editors. (1993). *Talking Data: Transcription and Coding in Discourse Research*. Lawrence Erlbaum, Hillsdale.

Fast, K., Fujiwara, E., and Markowitsch, H. (2006). *Bielefelder Autobiographisches Gedächtnis Inventar (BAGI)*. Swets & Zeitlinger, Lisse.

Fast, K., Fujiwara, E., Schröder, J., and Markowitsch, H. (2007). *Erweitertes Autobiographisches Gedächtnis Inventar (E-AGI)*. Harcourt, Frankfurt a. Main.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

Hansen, J. H., Deller, J., and Seadle, M. (2001). Engineering challenges in the creation of a National Gallery of the Spoken Word: Transcript-free search of audio archives. In *Proc. IEEE ACM Joint Conf. Digital Libraries*, pages 235–236.

Härting, C., Markowitsch, H. J., Neufeld, H., Calabrese, P., Deisinger, K., and Kessler, J. (2000). *Wechsler Gedächtnis Test – Revidierte Fassung*. Huber, Göttingen.

Hazen, T. J. (2006). Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings. In *INTERSPEECH 2006 – 7th Annual Conference of the International Speech Communication Association*.

Horn, W. (1983). *Leistungsprüfsystem (LPS): Handanweisung*. Hogrefe, Göttingen, 2. revised and improved edition.

Juster, F. T. and Suzman, R. (1995). An overview of the health and retirement study. *The Journal of Human Resources*, 30:7–56.

Kim, C. and Stern, R. M. (2008). Robust Signal-to-Noise Ratio Estimation based on Waveform Amplitude Distribution Analysis. In *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association*, pages 2598–2601.

Lehr, U., Thomae, H., Schmitt, M., and Minnemann, E. (2000). Interdisziplinäre Längsschnittstudie des Erwachsenenalters: Geschichte, theoretische Begründung und ausgewählte Ergebnisse des 1. Messzeitpunktes. In Peter Martin, et al., editors, *Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, pages 1–16. Steinkopff.

Levy, R. (1994). Aging-associated cognitive decline. *International Psychogeriatrics*, 6(01):63–68.

Martin, P. and Martin, M. (2000). Design und Methodik der Interdisziplinären Längsschnittstudie des Erwachsenenalters. In Peter Martin, et al., editors, *Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, pages 17–27. Steinkopff.

Martin, P., Grünendahl, M., and Schmitt, M. (2000). Persönlichkeit, kognitive Leistungsfähigkeit und Gesundheit in Ost und West: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE). *Zeitschrift für Gerontologie und Geriatrie*, 33(2):111–123.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–939.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L., and Cernocky, J. H. (2011). RNNLM - Recurrent Neural Network Language Modeling Toolkit. In *IEEE Automatic Speech Recognition and Understanding Workshop*.

Moreno, P. J., Joerg, C. F., Van Thong, J.-M., and Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *International Conference on Spoken Language Processing, ICSLP*, pages 2711–2714.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., and Fillenbaum, G. (1989). The consortium to establish a registry for Alzheimer's disease (CERAD) Part I: Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, 39:1159–65.

Nouza, J., Cerva, P., Zdansky, J., Blavka, K., Bohac, M., Silovsky, J., Chaloupka, J., Kucharova, M., Seps, L., Malek, J., and Rott, M. (2014). Speech-to-Text Technology to Transcribe and Disclose 100,000+ Hours of Bilingual Documents from Historical Czech and Czechoslovak Radio Archive. In *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, pages 964–968.

Oswald, W. and Fleischmann, U. (1993). *Das Nürnberger Alters-Inventar NAI. Kurzbeschreibung, Testanweisung, Normwerte, Testmaterial*. Hogrefe, Göttingen.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Prince, M., Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., and Prina, M. (2015). *World Alzheimer Report 2015. The Global Impact of Dementia: an Analysis of Prevalence, Incidence, Cost and Trends*. Alzheimer's Disease International, London.

Román, G. C., Tatemichi, T. K., Erkinjuntti, T., Cummings, J., Masdeu, J., Garcia, J. a., Amaducci, L., Orgogozo, J.-M., Brun, A., Hofman, A., et al. (1993). Vascular dementia diagnostic criteria for research studies: Report of the NINDS-AIREN International Workshop. *Neurology*, 43(2):250–250.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2):81–88.

Schönknecht, P., Pantel, J., Kruse, A., and Schröder, J. (2005). Prevalence and Natural Course of Aging-Associated Cognitive Decline in a Population-Based Sample of Young-Old Subjects. *American Journal of Psychiatry*, 162(11):2071–2077.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). GlobalPhone: A Multilingual Text & Speech Database in 20 Languages. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., and Uhmann, S. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 10:353–402.

Statistische Ämter des Bundes und der Länder. (2011). Bevölkerungs- und Haushaltsentwicklung im Bund und in den Ländern. *Demografischer Wandel in Deutschland*, 1.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *International Confonference on Spoken Language Processing*.

Telaar, D., Wand, M., Gehrig, D., Putze, F., Amma, C., Heger, D., Vu, N. T., Erhardt, M., Schlippe, T., Janke, M., Herff, C., and Schultz, T. (2014). BioKIT - Real-time decoder for biosignal processing. In *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, pages 2650–2654.

Tewes, U. (1991). *Hamburg-Wechsler-Intelligenztest für Erwachsene: Handbuch und Testanweisung*. Huber, Göttingen.

Veselý, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association*, pages 2345–2349.

Wendelstein, B. and Sattler, C. (2011). Das ILSE-Korpus. Eine korpuslinguistische Perspektive psychologisch-psychiatrischer Forschung am Beispiel der Alzheimer-Demenz. In Ekkehard Felder, et al., editors, *Korpus-*

*pragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*, pages 488–511. De Gruyter, Berlin/Boston.

Wendelstein, B. (2016). *Gesprochene Sprache im Vorfeld der Alzheimer-Demenz. Linguistische Analysen im Verlauf von präklinischen Stadien bis zur leichten Demenz.* Winter, Heidelberg.

World Health Organization and Alzheimer's Disease International. (2012). *Dementia: a public health priority*. World Health Organization.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.

## 11. Language Resource References

Schultz, Tanja. (2014). *GlobalPhone German*. distributed by ELRA, GlobalPhone, ISLRN 937-733-002-847-8.