Top-down processing of words in degraded speech

Olsen, Ludvig and Larsen, Kristian, 2016, Cognition and Communication
Supervised by Tylén, Kristian

1.1 Abstract (Ludvig and Kristian)

It is genuinely accepted within the field of cognitive hearing science that language comprehension emanates from an interaction between several neural processes. This study examines the interplay between bottom-up and top-down processing. The conducted experiment was designed to gather information, particularly on the role of the higher-function processes in comprehension of degraded speech, using a staircasing design. A novel audio degradation technique, Sample Silencing, was used to degrade speech stimuli, respectively words that are highly frequent and words that are less frequent in the Danish language, along with pseudowords resembling Danish phonology. Different types of investigative models have been included in the analysis of our data counting several tables and visual representations.
Top down processing has been found to influence the comprehension of degraded speech, though it is difficult to establish to which extend the frequency of a word matters in this process. Real words have been found significantly easier to comprehend than pseudowords.

Keywords will include: bottom-up, top-down, working memory, auditory pathways, Wernicke's area, Broca's area, gamma band activity, speech degradation, Sample Silencing

2.1 Introduction (Kristian)

Animals use sound to signal one another when danger is near or to express their location. Humans hear the wind blow, the dogs' bark and the yell of their neighbor. However, humans differ from animals in that they utter and understand highly complex sounds such as languages. In the field of auditory cognitive science, researchers try to answer questions on how we percept such multifaceted sounds as language. This research emerged as a result of technological development, allowing researchers to perform digital tasks that study the interplay between language input and cognition. Studies have been conducted carrying out experiments on how language is processed under challenging conditions for the listener e.g. using white noise or degraded speech (Hunt, M.J., Lefdbvre, C., 1989), (Hannemann, R., et al. 2007). One of many indeterminate areas of investigation is how exactly *meaning* is created from sound information. Early studies suggested that higher-order cognitive functions were independent and therefore had no impact on the slower processes of perception (Arlinger S. et al., 2009, p. 371-384.) Current research in the field of cognitive hearing science seeks to understand the interaction of brain processes. It is believed that the future holds a more distinct understanding of the interaction between bottom-up and top-down processing (Rönnberg J., et al. 2010, p. 1-2).

In everyday life, humans communicate under different, sometimes horrible acoustic conditions; and in these environments people usually manage to comprehend one another. A neuroimaging study from 2003 explores how the frontal areas of the brain are activated alongside the temporal areas when processing language. This study stated that the frontal area holds the higher-level cognitive functions such as top-down processing and the temporal areas contain the lower-level functions like bottom-up processing. Following on from the common saying that states: neurons that fire together wire together, it may be a reasonable assumption that language understanding in the brain involves a highly connected neural network (Scott, S.K., Johnsrude, I.S., 2003). Specifically when communicating in terrible acoustics, or when exposed to degraded speech, the influence of the mechanisms that are considered to be higher-function is thought to be highly operative. J. Obleser *et al.* have argued that semantic context plays a significant role in speech comprehension. Their study on how a verbal signal interacts with the semantic expectancies displayed the interplay between bottom-up and top-down processes in understanding speech. Degraded speech signals increased the influence of top-down processing on comprehension within the semantic frame of the sentence. Thus they conclude that semantic connections are important, though the hierarchy of bottom-up acoustic processing and top-down semantic understanding is not yet fully explained (Obleser, J., Kotz, S.A., 2011, p. 713–723).

Our study examines the interplay between bottom-up and top-down processing in degraded speech without a dialogue specific context. The first part will describe the theoretical foundation of the

project including the definition of relevant terms in the research area of the investigation and a technical review of speech degradation techniques. This will be followed by a systematic walkthrough of the structure of the experiment. The data obtained from the experiment will be analyzed with the aim of clarifying/disqualifying the hypothesis. Using the relevant terminology, the discussion will debate how well the results interact and support other studies on the matter, and how the results contribute to the area of auditory cognition.

- We hypothesize that participants will recognize the high frequency stimuli words before the low frequency words, and both the high frequency and low frequency words before the pseudowords.

2.2 Theoretical foundation (Kristian)

Bottom-up processing is best described as a process driven exclusively by the stimulus itself. Hence only the stimulus itself impacts the perception. This process is considered one of the low-level processes in auditory cognition. Using fMRI scans of the brain, a study by Binder et al. (2000) suggested that the posterior temporal area and dorsolateral temporal areas in the brain are very active when exposed to all kinds of auditory stimuli (Binder et al., 2000, p. 524); Hence the implicit filtering of all kinds of sound material initiates as a bottom-up process. Sophisticated sounds, such as language, are processed differently from basic sounds like noise or tones. The anterior part of the temporal region shows strong activation when processing language (Binder et al., 2000, p. 521). This part of the temporal area is linked to the ventrolateral prefrontal cortex, which is connected to the higher-level cognitive functions that perform the top-down processing of words.

Top-down processing is driven by background knowledge. This creates expectations when hearing sounds based on previous understanding and experience. It is considered a higher-order cognitive process as it primarily occurs within interactions with frontal lobe functions including the working memory (WM) (Arlinger, S. et al. 2009, p. 371-384). The interplay of sensory input and the anticipations driven by the frontal-lobe systems, e.g. working memory, performs the function of recognizing and creating meaning in language perception under challenging conditions (Hannemann, R, et al. 2007, p. 139).

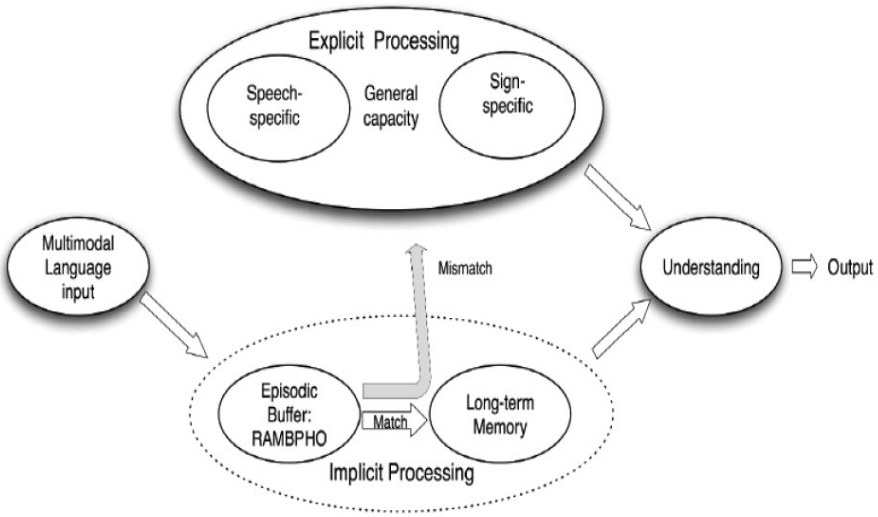

Figure 1: A working memory system for Ease of Language Understanding,
Adopted from Rönnberg, Jerker, et al., 2010, p. 265

In a study by Rønnberg et al. (2010) they detail the role of working memory in the Ease of Language Understanding model (Rönnberg, Jerker et al., 2010, p. 265)(See: **Figure 1**). The first part '*multimodal input'* refers to any kind of multi-sensory linguistic information. As previously described, language perception is considered to be a higher-order cognitive function overall involving both high-level and low-level cognitive processes. In this model the initial processing of language is explained as *"RAMBPHO, Rapidly, Automatically, and Multi-modally Bound together to form Phonological streams of information"* (Rönnberg, Jerker et al., 2010, p. 265). If the phonological input is comprehensible, the working memory implicitly matches the input with stored material in the Long-Term Memory (LTM), creating the understanding. If the speech is degraded, or in any way compromised, the implicit processing can fail to match the input to the long-term memory; which is when a mismatch happens. In those situations it is suggested that other language factors such as the semantic context and/or specific characteristics in the dialogue may compensate for the lack of information and help the phonological processing so that even compromised speech may be understood through a top-down process (Rönnberg, Jerker et al., 2010, p. 263-269). Hence the model explains the explicit processing that occurs when mismatches happen. This means that a WM process occurs both when examining the sound momentarily (implicit processing) and when analyzing the sound retrospectively to counteract the incompatibility between the sound input and the LTM (explicit processing). Thus the working memory is involved in both top-down and bottom-up processes. We want to highlight the distinction between a cognitive function and a cognitive process. As previously mentioned the bottom-up process is considered low-level and the top-down process is considered a high-level process. However, both processes include higher-order cognitive functions of the frontal

regions, and both processes initiate as lower-level sensory inputs, which activates the auditory areas in the temporal lobe; hence it is also relevant to comment on the brain anatomy of the auditory system and the difference between receptive and expressive areas. When sound is received the Wernicke's area is considered the receptive area that facilitates understanding; thus the neural pathways of comprehension initiates here. When expressing language the Broca's area is considered the expressive area that mediates the communication as a multimodal output (speaking, writing, gesturing). Wernicke's area is located in the posterior left region of the temporal lobe, and as previously described it is highly active when exposed to all sound input. The Broca's area is located anterior to the primary auditory cortex and is connected to the ventrolateral prefrontal cortex, the part that is found to be highly active in top-down processing. We want to emphasize that auditory perception is not fully understood, and all mentioned processes play a role in a yet unexplained neural network (Friedenberg, J., Silverman, G., 2011, p. 294-304).

2.3 Stimulus and memory matching (Ludvig)

A modulation of gamma band activity (GBA) has been observed in both the auditory and visual domain when participants recognize certain stimuli as opposed to new stimuli similar in complexity (eg. Lutzenberger, W. et al., 1994, p. 115, 117; Herrmann, C.S. et al., 2004b, p. 1-5; Lenz, D. et al., 2006, p. 31-37; Hannemann, R. et al., 2007). Concluding on an EEG experiment, that found this kind of increase in GBA in the 40Hz range in the left hemisphere when participants heard degraded speech stimuli that had previously been presented in its non-degraded form, R. Hannemann et al. (2007) linked this observed enhancement to a process of matching the degraded sensory input to top-down lexical memory traces.

Christoph S. Herrmann et al. (2004a) describe this matching process between stimulus-related information and memory contents in their "match-and-utilization-model" (MUM). In this model the matching result - either *match* or *mismatch* - is utilized by updating memory content, changing behaviors or reallocating attention. When a positive match occurs this model predicts an enhanced early gamma response, whereas a late gamma response is hypothesized a sign of utilization.

In an experiment using pitch continuation in a series of sinusoidal tones to create, and at times violate, expectations Jeanette Schadow et al. (2008) found an increase in early GBA when expectations were met, i.e. the expected tone was played, indicating that an increase in early GBA might in a broader sense be linked to a top-down prediction and not only a matching of existing memory content.

2.4 Speech degradation techniques (Ludvig)

Multiple audio degradation techniques have been utilized on speech stimuli in past experiments. An example of an often-used additive approach is to add noise alongside the signal. This allows for experimentation with signal-to-noise ratios finding the ratios of possible/impossible comprehension. Disadvantages with using noise are, that the noise itself might lead to delayed and attenuated brain responses; that frequencies are added to the signal; and that the noise might be interpreted as a different sound source than the speech signal (Miettinen, I. et al., 2011, p. 298-299; Miettinen, I. et al., 2010, p. 2). Other approaches could be to alter and/or remove audio information such as frequencies (spectral processing), amplitude, e.g. reduction of amplitude resolution, or in the time domain(temporal processing), e.g. Phonemes, etc. (Krishnamoorthy, P., 2009, p. 137; Miettinen, I. et al., 2010, p. 3-4)
Choice of degradation method can be motivated by the aim to simulate or model language deficits or impairments (Dick, F., 2003, p. 535)

3.1 Experimental design (Ludvig)

This section will cover the technical details about the conducted experiment. Selected stimuli will be presented with a thorough explanation of how it was recorded and digitally processed. Finally we will review the experimental process.

3.2 Stimuli (Ludvig)

The stimuli consist of **24** 3-syllable words, that have been recorded and gradually degraded using our own digital, degradation algorithm (see: **3.4**). The words are divided into 3 categories: **8** Danish words that are highly frequent in the Danish language according to our chosen corpora. **8** words that are less frequent in the Danish Language, and **8** pseudowords that we have made ourselves using syllables from the Danish Language.
We have tried to avoid compound words as these can be made up by more or less frequent words. A few pseudowords, such as “erkaplads”, can be considered a compound word though, as part of the word, “plads”, is actually a common Danish word itself. This was recognized post-hoc.

High frequency words have been chosen by looking up words, assumed to be highly frequent, in the corpus KorpusDK, selecting words most frequently appearing in the corpus. The same procedure has been used to select less frequent words, though choosing the words appearing the least, or not at all, in the corpus.

KorpusDK[1] (2007 version) is a collection of two corpora, **Korpus 90**, gathered in the years 1983-1992, containing texts from newspapers, magazines, books, and more.[2] The other, **Korpus 2000**, was gathered in the years 1998-2002 and contains texts from newspapers, magazines, books, schools, unions, companies, websites, and individuals. [3]

In this corpus inflected forms have been included in our search as it reflects the use of the word in different contexts.

The appearance frequencies of the 16 stimuli words in KorpusDK have been compared to the appearance frequencies in the two corpora **Information(1998-2008)** and **Wikipedia** found on http://corp.hum.sdu.dk/cqp.html[4], not including inflected forms in searches.

These two corpora confirm the gap between our highly frequent words and our less frequent words.

Table 1: Word stimuli (High frequency and low frequency)

Word	**DK2007**	**Information (1998-2008)**	**Wikipedia**
High frequency			
arbejde	56911	46699	1098
menneske	47965	11001	312
begynder	32407	12198	400
mulighed	30658	21455	544
billede	20914	12260	351
resultat	17099	7032	229
politi	15892	6424	81
lejlighed	10821	6147	160

[1] KorpusDK: http://ordnet.dk/korpusdk_en/facts/available-corpora?set_language=en
[2] Korpus 90: http://ordnet.dk/korpusdk_en/facts/available-corpora/suppliers-of-texts-to-korpus-90?set_language=en
[3] Korpus 2000: http://ordnet.dk/korpusdk_en/facts/available-corpora/suppliers-of-texts-to-korpus-2000?set_language=en
[4] Information(1998-2008) & Wikipedia corpora: http://corp.hum.sdu.dk/cqp.html

Low frequency			
konstabel	90	29	4
tematik	71	218	1
viagra	60	215	1
konsonant	50	5	17
retledning	14	8	0
besejring	4	14	1
omnibus	3	13	3
historik	3	36	3

Our pseudowords have been created with a python generator that randomly puts three Danish syllables together. 14 pseudowords sounding like possible Danish words were tested in Google searches to avoid company names etc. before being recorded. The 8 best-recorded pseudowords were then selected.

Table 2: Word Stimuli (Pseudowords)			
banekop	gebafon	erkaplads	opdande
ekspimo	onopal	galiti	mentinyt

3.3 Recording (Ludvig)

All words have been spoken in a randomly shuffled order to avoid possible effects from vocal fatigue. All words have been recorded 4 times for our speaker to get some practise on the pseudowords in order to minimize the effect of words not previously pronounced.
A steady tempo and clear articulation has been attempted to ease the spelling of the word when hearing it once in its non-degraded form. This has only been tested by ourselves while choosing the best recordings from the four takes to make it as easy as possible to hear the word. Later analysis of

the audio files indicated that pseudowords might have been spoken a bit slower than high frequency words (**See: 5.3 Analysis of Stimuli**)

Sounds have been recorded as *.wav* at *24bit*, *44.1kHz* through the following hardware chain:
Microphone: SE Gemini II; preamp: Great River ME-1NV; interface: Focusrite Liquid Saffire 56;
DAW: Logic Pro X
Cut up sounds have then been exported as *.wav* at *16bit*, *44.1kHz*.

In total we have 264 stimuli audio files with a mean length of 0.789s, SD = 0.1s.

3.4 Audio degradation (Ludvig)

Normal CD quality audio files contain 44,100 samples per second. This is called the sample rate.[5] This means that for every 44,100th of a second we have a value describing the analog audio at that moment in time. If this value is set to 0, there is silence. So by setting more and more samples to 0, we remove more and more information from the audio file.

This technique, that we call **Sample Silencing**, has been used to gradually(in layers) remove audio information from our audio files like this:

Table 3: Kept samples

Layer (0-5)	Kept samples	Layer (6-10)	Kept samples
0	1/100	6	60/100
1	10/100	7	70/100
2	20/100	8	80/100
3	30/100	9	90/100
4	40/100	10	100/100
5	50/100		

[5] Sample Rate: https://en.wikipedia.org/wiki/Sampling_%28signal_processing%29#Sampling_rate

This means that in layer 4, we have told our script to keep 40 values every 100th sample and set the last 60 to a value of 0. This goes on, until reaching the end of the audio file (See: **Figure 2**)

These values have been chosen because they create the wanted auditory effect for our experiment, i.e. a transition from no recognisability in the first layer (baseline), possible recognisability in the mid layers (4-6), and certain recognisability in the end.

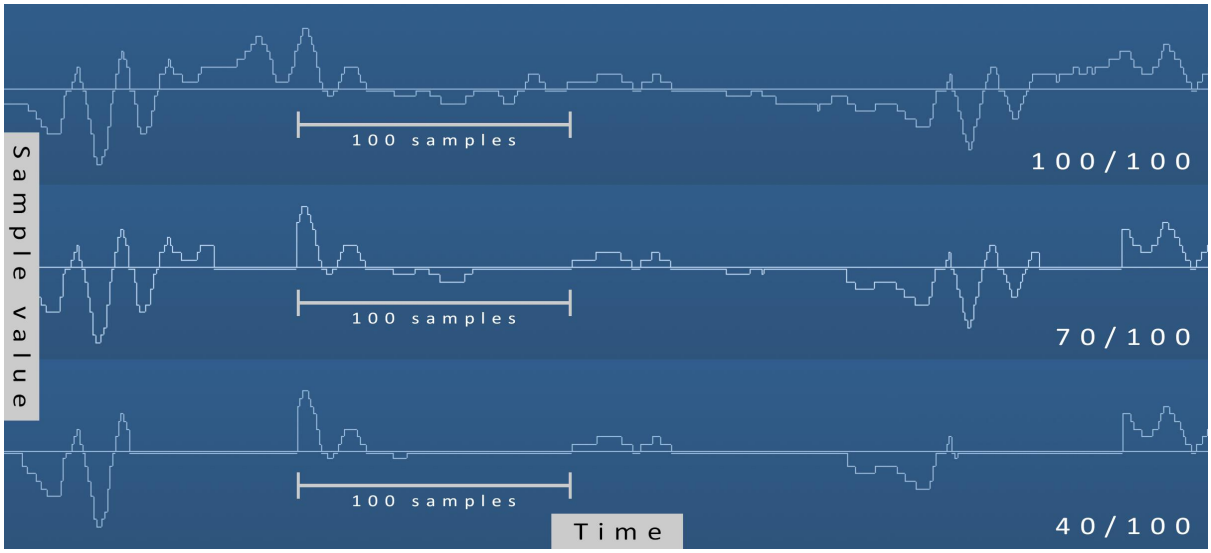

Figure 2: Comparing three layers from the stimuli sound "Historik"
From the top: layer 10, layer 7, and layer 4. This is a zoomed in visualization of the sounds from the digital audio station Logic Pro X[6]. The straight lines represent the sample value 0 in each sound. Samples above the line have positive values, while samples below the line have negative values. Note that Logic puts 0-values right below the line in its visualization.
We see how the sound files share the same *sample values(Y)* in the beginning, *time(X)*, of the 100 samples but one by one drop to the value 0.

When exporting our degraded sound files, file lengths from layer 0 to layer 10 varied gradually by up to 2.17ms. Aligning them manually assured us this had not affected the audio content, i.e. it had no effect on our experiment but should be noted if using the stimuli for more time-sensitive experiments in the future.

[6] Apple Logic Pro: http://www.apple.com/dk/logic-pro/

3.5 Experimental process (Ludvig)

After giving informed consent participants were briefed on how to carry out the experiment. This info was also presented in a written form to all participants to avoid misunderstandings from the researcher's briefing. Many participants asked the present researcher questions even while conducting the experiment and got answers allowing them to feel comfortable with continuing without revealing compromising details about the experiment.

Briefing points presented as text were(own translation from Danish):

- You are about to listen to 24 different words
- No words with “æøå”(Danish letters) will be included.
- There will be both real and made up words.
- Your job is to guess the words as quickly as possible.
- Words will gradually become easier to understand.
- If you cannot hear what is being said, then try to guess. If you have
- Absolutely no idea you can type "n" instead of typing the word.
- In the beginning it will be very hard to hear the words, but it is important that you listen carefully.
- It is more important that you guess early on than spell correctly.
- If you experience a word that you have already typed, you have likely either misheard or misspelled the word last time. Try again.

Participants were placed in front of a computer in relatively quiet rooms without big, obvious, external disturbances. They were then given a set of headphones, respectively Beyerdynamic DT770 Pro (researcher: Ludvig) and Samsung Level Over (researcher: Kristian).
A python script run in PsychoPy version 1.82.1[7] presented a dialogue box for gathering personal information such as name, age, gender, years of school, and a selection of the researcher having briefed the participant (including headphone model). Then the written briefing appeared for the participant to read.
After this the experiment began: The participant heard an audio file once and had to type either "n" (i.e. “I don't even have a guess”) or the word heard/guessed upon. Then the participant pressed "Enter" and a blank page appeared along with the next audio file.

[7] PsychoPy: http://www.psychopy.org

Participants were presented to the stimuli in a within-participant, staircasing design. First they heard all words in layer 0 (1/100 samples) in a shuffled order as a baseline; then all words in layer 1 (10/100 samples) in a shuffled order, etc.
Once a word had been spelled correctly it stopped appearing. This rule resulted in a mistake, where some participants would answer "politi" at the word "galiti", meaning that the word "politi" would never actually be guessed as it stopped appearing. This was handled in the analysis, hence not having any big consequences on the results.
Once all words had been written correctly or the participant had gone through all 11 layers (0-10), the experiment stopped. The participant was then asked about the experience and important things were noted down.

The experiment was conducted in Danish. Besides participants' responses, we collected reaction times measured from heard stimuli to first typed character.

4.1 Participants (Ludvig)

32 subjects (19 females; 13 males; aged 15-55 years, mean age: 27.91 years, standard deviation SD=13.2 years; 9 - 17 years of education, mean: 13.31, standard deviation SD=1.6) participated in our experiment. All were native Danish speakers.
We obtained informed consent from all participants.

5.1 Analysis (Ludvig)

In this section we will cover the handling of experimental data; coding of certain errors made by participants, statistical models, and choosing data for result tables.

5.2 Coding (Ludvig)

In some cases participants made spelling mistakes though obviously having heard the right word. As our hypothesis mainly concerns hearing and not spelling, we decided to include these answers as correct and hence coded for these mistakes.
Only when pronunciation of the participant's spelling was considered to be identical to the original word, was it coded for. Examples of common corrected words were "expimo" ("x" and "eks" sounds the same in Danish) and "opdante" (often "t" and "d" are pronounced the same in Danish). (See **Table 4**)

Using python we extracted the layers (0-10) where participants first answered the word correctly or had been corrected in the coding process. When a participant missed a word altogether the word was set to "layer 11" in order to count these words in when calculating means etc. (See **Table 4**)
Because of a mistake in the experiment script, when people answered "politi" at the stimulus "galiti", before answering "politi" at the stimulus "politi", they would not face the word "politi" again, as the script regarded it already correctly answered. When this happened we did not set the layer to anything, hence the word "politi" in these few(5) instances was not counted in when calculating mean etc.

5.3 Analysis of stimuli (Ludvig)

Testing for confounding factors, we performed an ANOVA on the length of our audio files between word types and found a significant difference (**F(2,21)** = **6.818**, p = **.005**). A post-hoc Tukey test showed that high frequency words were significantly shorter than low frequency words by **0.12 seconds** ± **0.04234**(standard errors), p=**0.02**, and that high frequency words were significantly shorter than pseudowords by **0.14** seconds ± **0.04234**(standard errors), p=**0.007**. There were no significant length difference between low frequency words and pseudowords (p=**0.87**).
Considering the mean audio file length of 0.789s, SD = 0.1s, this could have an effect on the perception of the words. Hence we included it in our complex linear mixed effects model(**5.4**).

5.4 Statistical models (Ludvig)

In order to find the relationship between **Word Types** (high frequency, low frequency, and pseudo) and the **layers** in which participants answered words correctly; we used R (R Core Team, 2015) and *lme4* (Bates, D., Maechler, M., Bolker, B., and Walker, S., 2015) to perform two linear mixed effects analyses.

First off we made a simple model (M1) checking the difference between word types for each participant. As fixed effects, we entered **word type** into our model. As random effects, we had intercepts for **subjects**.

Then, to control for various confounding factors, we made a more complex model(M2) taking differences between items(words), researcher(including headphones), and lengths of the audio files into account. As fixed effects, we entered **word type** and **audio file length** (without interaction term) into our model. As random effects, we had intercepts for **subjects**, **items**, and which **researcher** conducted the experiment (including headphone type).
A visual inspection of residual qq plots did not reveal any obvious deviations from normality in either model, while Levene's test found homoscedasticity. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.

Post hoc tukey tests were made using *multcomp*(Hothorn, T., Bretz, F., and Westfall, P., 2008) in order to find the relationships between the various word types.

To test whether participants heard individual words at the same layer, we calculated **mean** and **standard deviation** for each word (**Table 4**). We then performed an ANOVA on these **Standard Deviations**, predicted by **word type**, to see, whether participants agreed on the layers at which each stimulus became comprehensible, across word types.
Using the by() function in R (R Core Team, 2015), we affirmed normal distributions, while Levene's test found homoscedasticity.

5.5 Data for tables (Ludvig)

We have extracted relevant information from our data to support a discussion of top-down and bottom-up processing in word recognition.
In table 5 and 6 (**6.2 Most common wrong guesses**) we have focused on the most common wrong guesses made by the participants as we suspected these to be highly frequent in the corpus KorpusDK. Words have only been gathered once from each participant, meaning that any duplicate answers by a participant are not included in the tables.
As some guesses were compound words, we checked both the word in its entire form and relevant word parts (eg. "markedsplads", "marked", and "plads").

6.1 Results (Ludvig and Kristian)

In our simple model (M1), word types affected in which layer participants answered the right word **(χ^2(2)=500.98, p<2.2e-16)**. A post hoc Tukey test showed that **high frequency words** and **pseudowords** differed significantly with pseudowords being answered later by **3.70 layers ± 0.1467(standard errors)**, **p<1e-05**, and that **low frequency words** and **pseudowords** differed significantly with pseudowords being answered later by **2.97 layers ± 0.1460(standard errors), p<1e-05**, and that **high frequency words** and **low frequency words** differed significantly with low frequency words being answered later by **0.72 layers ± 0.1469(standard errors)**, **p<1e-05**.

In our complex model (M2), word types affected in which layer participants answered the right word **(χ^2(2)=24.84, p=4.038e-06)**. A post hoc Tukey test showed that **high frequency words** and **pseudowords** differed significantly with pseudowords being answered later by **3.80 layers ± 0.6656(standard errors)**, **p<1e-04**, and that **low frequency words** and **pseudowords** differed significantly with pseudowords being answered later by **2.99 layers ± 0.5367(standard errors),**

p<1e-04. There was no significant difference between high and low frequency words, **0.80 layers ± 0.6328(standard errors)**, **p=0.408**.

There was a significant difference in standard deviations between word types(**F(2, 21)** = **4.695**, p=**.021**). A post-hoc Tukey test showed that high frequency words had significantly smaller standard deviations than pseudowords by **0.58 layers** ± **0.1972**(standard errors), p=**0.02**, while no significant difference were found between high frequency words and low frequency words (**0.44 layers** ± **0.1972**(standard errors), p=**.0866**), or between low frequency words and pseudowords (**0.14 layers** ± **0.1972**(standard errors), p=**.7749**).

Summary of responses:

Table 4: Mean, corrections and how many times a word was not found at all - per word

Word in word type	**Mean**	**SD**	**Corrections**	**Not Found At All**
High Frequency				
arbejde	4.16	0.68	1	0
begynder	4.47	0.62	0	0
billede	5	1.44	3	1
lejlighed	4.5	0.92	0	0
menneske	4.03	0.82	4	0
mulighed	5.53	0.98	1	0
politi	4.04	0.81	1	0
resultat	3.94	1.19	0	0
Low Frequency				
besejring	5.56	1.11	0	0
konsonant	4.94	1.52	1	0
konstabel	4.47	0.88	1	0

omnibus	5.47	1.61	0	0
retledning	7.78	2.07	2	3
tematik	4.75	1.74	0	1
viagra	4.94	1.24	0	0
historik	3.66	0.83	2	0
Pseudoword				
banekop	5.78	1.5	4	0
ekspimo	9.72	0.92	5	3
erkaplads	9.22	1.81	1	15
galiti	7.03	1.49	0	0
gebafon	8.88	2.25	0	14
onopal	9.88	0.91	0	9
opdande	7.61	1.78	22	2
mentinyt	7.28	1.42	0	0

6.2 Most common wrong guesses (Ludvig and Kristian)

Table 5: The 5 most common wrong guesses(word) per stimulus word and a count of participants answering it(Freq.).

Ekspimo Word	ekspio	ekspilot	ekspilo	hbo	ekspedient
Ekspimo Freq.	18	11	10	6	5
Erkaplads Word	arkeplads	markedsplads	arkaplads	arbejdsplads	erkeplads
Erkaplads Freq.	15	14	13	8	8

Gebafon Word	mikrofon	gibafon	giberfon	gipafon	vibrafon
Gebafon Freq.	14	8	4	3	3
Onopal Word	unopal	homopal	omopal	envokal	olopal
Onopal Freq.	13	8	8	3	3
Opdande Word	opbande	opgande	akbande	bunke	opante
Opdande Freq.	13	4	2	2	2
Galiti Word	politi	matematik	politik	daliti	dementi
Galiti Freq.	12	4	4	3	3
Konsonant Word	konsulent	konfirmand	instrument	konsonent	aldrig
Konsonant Freq.	8	3	2	2	1
Tematik Word	synoptik	matematik	timatik	ansigt	semantik
Tematik Freq.	7	5	5	2	2
Mentinyt Word	mentimyt	internet	mentimet	mentinet	mentimit
Mentinyt Freq.	7	6	5	4	3
Banekop Word	edderkop	underkop	kaffekop	alekop	badekop
Banekop Freq.	6	4	3	2	2
Retledning Word	retning	rabledning	rapledning	drabledning	drapledning
Retledning Freq.	5	4	4	2	2
Besejring Word	besejling	besejre	eksamen	forsejling	adsac
Besejring Freq.	4	4	3	2	1
Omnibus Word	aarhus	underbuks	allesammen	besked	bonus
Omnibus Freq.	3	2	1	1	1
Arbejde Word	ordbog	banko	blabe	boom	brrr
Arbejde Freq.	2	1	1	1	1

Begynder Word	ideer	siger	agere	backamon	billede
Begynder Freq.	2	2	1	1	1
Billede Word	bog	gulerod	ur	banjo	bav
Billede Freq.	2	2	2	1	1
Menneske Word	hjemmesko	belzebub	endicso	enusko	ingenting
Menneske Freq.	2	1	1	1	1
Viagra Word	europa	mallorca	uaka	uoka	abe
Viagra Freq.	2	2	2	2	1
Lejlighed Word	dudadum	flimmer	lattergas	mandag	mavep
Lejlighed Freq.	1	1	1	1	1
Mulighed Word	alle	bbbr	bordben	det	englevoev
Mulighed Freq.	1	1	1	1	1
Politi Word	anders	anse	bo	diopsje	kantine
Politi Freq.	1	1	1	1	1
Resultat Word	bamse	citron	diskotek	eltog	forstod
Resultat Freq.	1	1	1	1	1
Konstabel Word	amstabel	andersfogh	badomski	computer	forstaet
Konstabel Freq.	1	1	1	1	1
Historik Word	allemand	drrstdum	estudent	himstergims	himstregims
Historik Freq.	1	1	1	1	1

Table 6: Most frequent overall wrong guesses and estimated frequency hereof in the Danish Language				
Wrong Guess	**By (n) Participants**	**Freq. KorpusDK**	**Freq. KorpusDK Relevant word parts**	
ekspio	18	0	433(eks)	91(pio)
arkeplads	15	0	17(arke)	22726(plads)
markedsplads	14	232	9338(marked)	22726(plads)
mikrofon	14	626	42(mikro)	4(fon)
arkaplads	13	0	4(arka)	22726(plads)
unopal	13	0	158(uno)	69(pal)
opbande	13	0	127646(op)	1113(bande)
politi	12	15892	1067(pol)	10244(ti)
ekspilot	11	1	433(eks)	974(pilot)
arbejdsplads	8	5188	56991(arbejde)	22726(plads)
erkeplads	8	0	0(erke)	22726(plads)
konsulent	8	1705	23(kon)	4(lent)
homopal	8	0	181(homo)	69(pal)
omopal	8	0	6(omo)	69(pal)
mentimyt	7	0	0(menti)	1(myt)
europa	7	12891	2798(euro)	208(pa)
edderkop	6	342	7(edder)	1609(kop)
internet	6	5017	321(inter)	4169(net)

6.3 Visualization of results

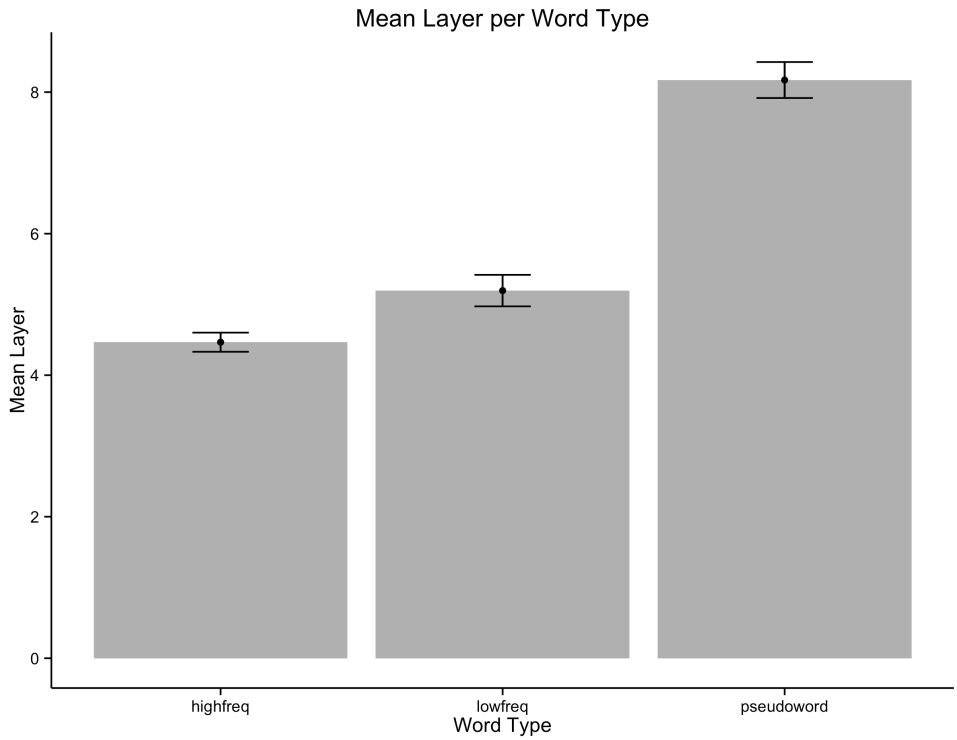

Figure 3: Mean layer answered correctly per word type

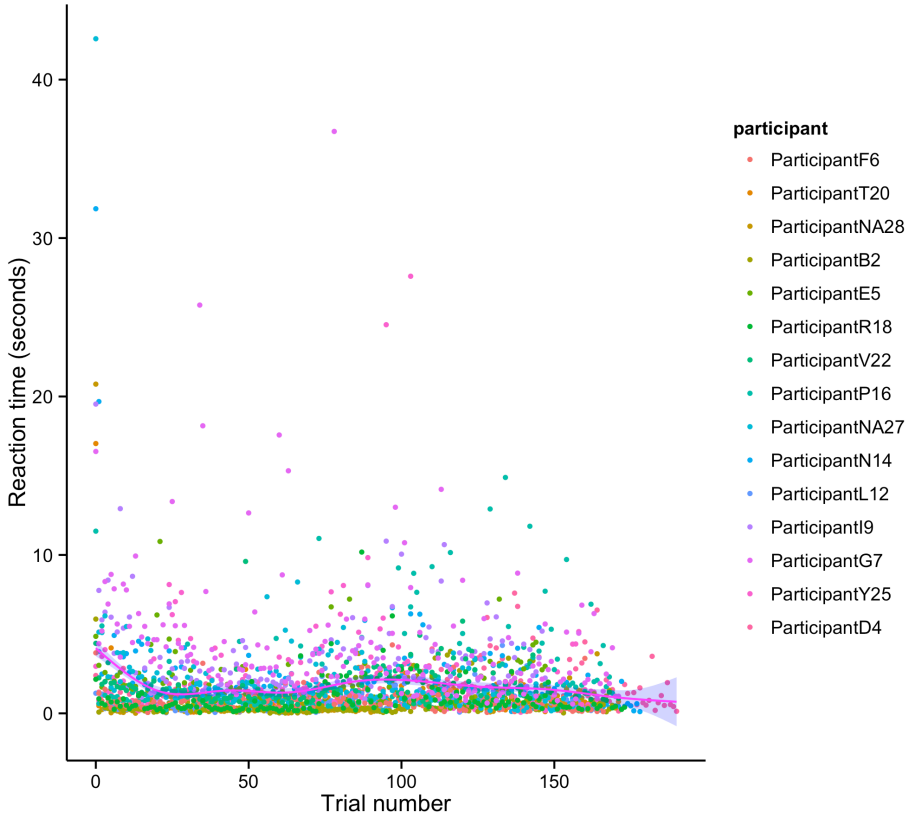

Figure 4: A selection of 15 anonymized participants showing development in reaction time (seconds) throughout the experiment (trials)

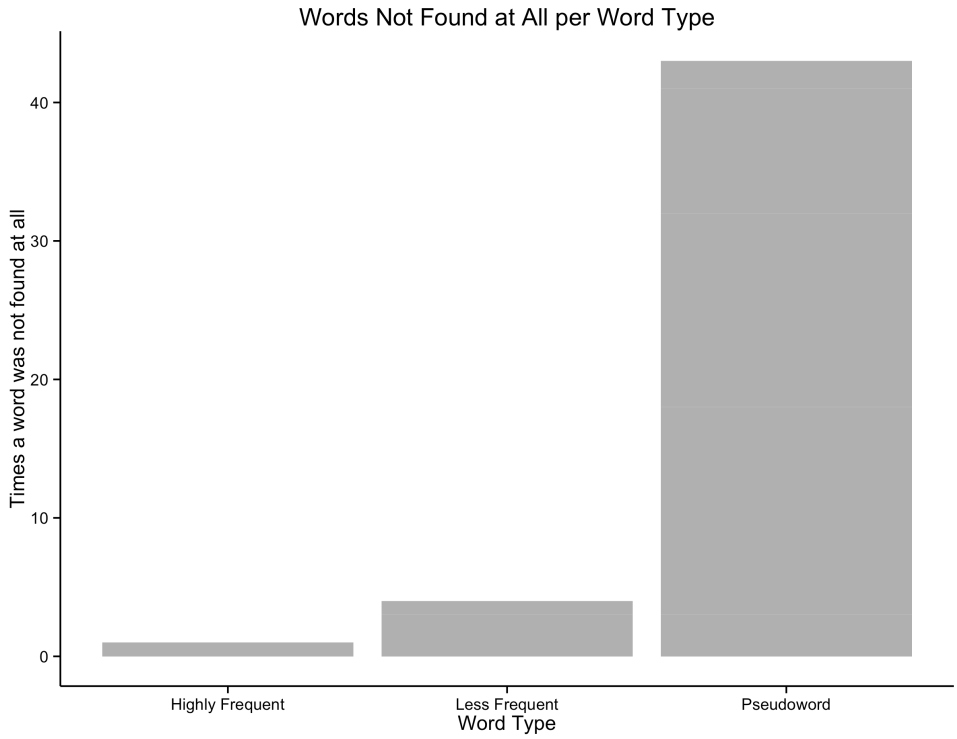

Figure 5: How many times a word was not answered correctly at all per word type.

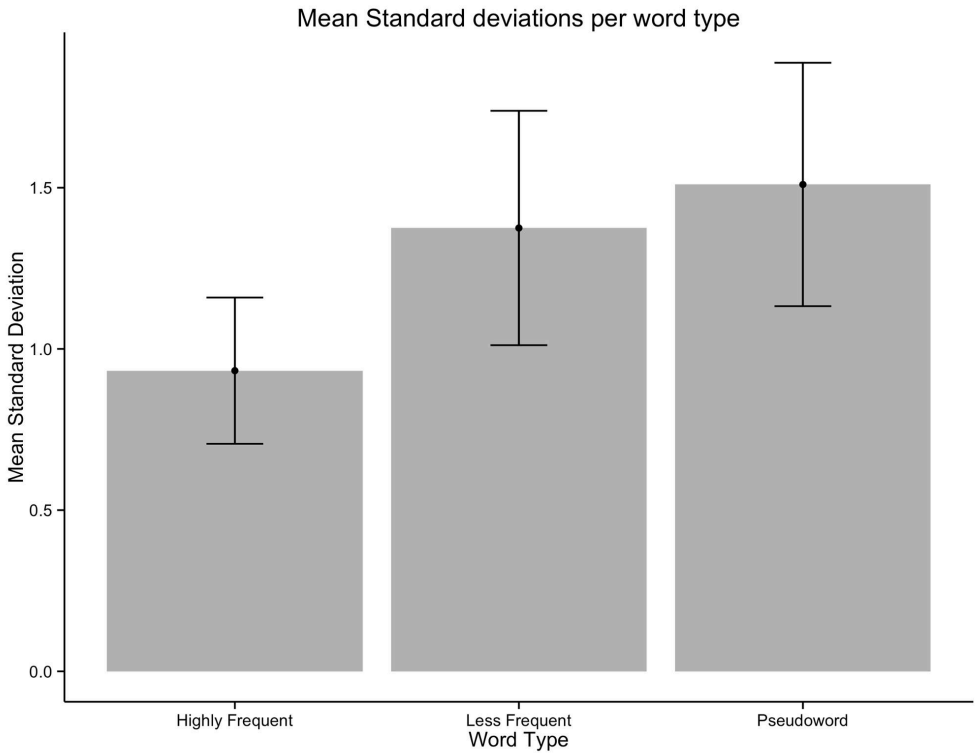

Figure 6: Mean standard deviation per word type

7.1 Discussion (Kristian)

In the discussion we will debate on how our analysis and results interconnects with the theory on the matter, mainly based on the complex model (M2) that contrasted our hypothesis to a certain degree. Particularly the role of top-down processing in our experiment will in focus. We will include a discussion of the wrong guesses based on the tables 5 and 6 in **6.2**.

From the analyses we found no significant difference between the high frequency words and the low words. We hypothesized a difference between the two. As mentioned in the theoretical chapter (p. 3-4), the top-down process is driven from background knowledge, i.e. within the semantic context and the dialog specific aspects of a conversation; and according to the MUM an affirmative word-matching with the long-term-memory content enhances the early gamma response; possibly making words that are frequently matched exceedingly responsive compared to infrequently used words. Correspondingly we thought high frequency words would be predictable and within a semantic context of day-to-day vocabulary; hence the top-down process would have wired the words according

to expectations, thus we expected that finding the more predictive high frequency words compared to the low words would be a quicker process. However, this was not the case in the complex model (M2). In the ease of Language Understanding model (Rönnberg, Jerker et al., 2010 P. 263-269) we saw how the implicit process matches a word with a lexical representation in the long-term memory. From the conducted experiment it seems that both high frequency and low frequency words were matched with a lexical representation in the long-term memory around the same time, 4th-5th layer. Thus it might have been possible to say that top-down processing only works context-specifically, and is therefore not an active part of comprehension when listening to words one by one. However, the very significant difference between pseudowords and regular words tells us something different. If all the words were simply hearable around layer 4-5, we should have found the recognition of pseudowords closer to regular words. Given that the pseudowords were created with Danish phonological syllables, the lexical representation should match the pseudowords approximately at the time where regular words were hearable. The pseudowords were generally assumed to be something else around the same time as the regular words were recognized. Pseudowords like 'erkaplads' and 'gebafon' were recognized as regular words like 'markedsplads' and 'mikrofon'. Participants were told to expect pseudowords so they wouldn't consciously try to force meaning into words that sounded peculiar. Given the fact that we searched for the layer where pseudowords became **hearable** we corrected some of the data-results post-hoc after realizing that the similarities between some Danish syllables e.g. 'te' and 'de' were extremely difficult to distinguish in the chosen pseudowords, even in layer 10. Nevertheless we found a significant difference, and taking into consideration that pseudowords were often misrecognized as regular words in the earlier layers, it appears as if top-down processing interferes comprehensively in the word retrieval. Apparently it is only when meaning is definitely not retrievable that we start relying on the raw material from the stimulus, creating a more bottom-up way of constructing meaning. This leaves us with two possible interpretations of our experiment.

1) Words are genuinely hearable around layer 4-5 making the WM implicitly match the words with the LTM (ELU); hence the top-down processing only affects the pseudowords, when trying to construct meaning from them.

2) Top-down processing interferes heavily as a reaction to degraded speech. This makes the idea of a 'genuine hear-ability' meaningless, as it is the interplay of processes that kicks in and affects both the regular words, which become recognizable early on, and the pseudowords, that initiates a process of constructing meaning and then later are accepted as nonsense.

The non-significant difference between high frequency and low frequency words could indicate that the lexical representation in the LTM between the two groups of regular words simply doesn't differ that much. In the theoretical part we mentioned how the WM process occurs both momentarily and retrospectively, we want to include another dimension saying that the WM process also occurs

prospectively. When matching words the WM takes part in the establishment of neural connections, as to why the lexical representations are retrievable in the first place. In that way the working memory contains a prospective dimension when generating the lexical representations necessary for future language perception. Post-hoc we ran some of the mistaken guesses on the pseudowords through the word corpus. Hereby we found that guesses attempting to find meaningful words, as opposed to pseudowords, were words of different, often high, frequencies, e.g. 'marked=9338', 'mikrofon=626', 'internet=5017'.

The extent to which top-down processing and the day-to-day vocabulary are linked is not determinable from our experiment. However, often the guesses on pseudowords were seeking meaning, indicating an expectancy-driven top-down process.

We suggest that the span of our lexical representations is highly comprehensive, and it would have been necessary to carry out the experiment with more than 24 words in order to significantly verify the differences between high frequency and low frequency words. In Figure 5 the count of words that weren't found at all is presented. When we looked into this matter it appeared that participants were very influenced by their own guesses, particularly with pseudowords. When participants spelled the pseudowords wrongly they often made very few if any alterations at all in later layers; even when more auditory information was presented they seemed unable to distinguish this version from the previous guesses, while other participants got the pseudowords stimuli right early on. This could indicate that the first guesses create some sort of lexical representation of the experienced pseudoword, making the participant predict the sound heard in later layers, why it becomes harder to hear the word correctly. There was also a slight trend between the high frequency and the low frequency, though not significant.

7.2 Evaluation of Sample Silencing (Ludvig)

Previous to this experiment various techniques have been used to degrade speech stimuli, each having advantages and disadvantages. New techniques might yield new research questions and are therefore relevant additions to scientists' toolboxes.

When using Sample Silencing, we are capable of maintaining the dynamics of the original stimuli. One researcher casually managed to recognize 4 stimuli words in the baseline layer(0), where only 1% (1/100 samples) of the audio information is present, followed by 13 in layer(1), indicating the actual experience and use of these dynamics.

Only information that was already in the original stimuli is present in the degraded stimuli as the degradation is caused by silencing samples in various sample intervals. By varying these intervals gradually, the technique offers great flexibility and control over the stimuli information. Though the effect of this method on human brain activity is unknown for now, we hypothesize that this kind of

stimuli might lead to cleaner results than what is seen from methods that alter, or add to, the audio information of the stimuli.
We recognize that this technique might not be suitable to simulate/model real world phenomena. This is not always necessary, though.

7.3 Evaluation of the experiment design (Ludvig)

Many participants uttered, on own initiative, that participation had been fun. In the beginning of the experiment many participants got frustrated, probably because they could not find any meaning in the stimuli of the first few layers, but once they were able to hear fragments of the stimuli, it became like a game. The staircasing design does contain an element of practice, and while participants knew that the task would become increasingly easier, it might have felt like they themselves were getting better, though not getting direct feedback on their performance.
We had a small script mistake, meaning that 5 participants did not get to guess 'politi'. This was handled by leaving out this word from those 5 participants' datasets, hence not having any big consequences on the results.

7.4 Conclusions (Ludvig and Kristian)

In line with recent studies that often denote the idea of a highly combined neural network (Friedenberg, J & Silverman, G, 2011, p. 294-304), we concur and suggest that our results display an interplay of bottom-up and top-down processing. The degraded speech seems to make the higher-order cognitive functions very influential on the comprehension; hence the top-down process affects all parts of our experiment. The two mentioned processes – bottom-up and top-down – have to be considered non-separable as they influence one another and are ongoing. Bottom-up doesn't seem to stop when top-down starts and vice versa.

7.5 Future research questions (Ludvig and Kristian)

- As the simple analysis model (M1) did indicate a difference between high frequency and low frequency words, it might be interesting to test a larger collection of words. Besides testing the words' frequencies in a corpus, a pre-investigation might be conducted gathering information of a different group of participants' associations to a word. A low frequency word as "viagra" might have more associations though being used less.
- Using high resolution, Sample Silenced speech stimuli, in controlled environments, it is possible to test speech comprehension differences between participants. This might be used to test whether twins raised apart differ the same as twins raised in the same environment or

even non-related participants of same age. If not so, it might yield new information on biological predisposal to language learning.

- In order to test whether participants create lexical representations of pseudowords in early layers, we could train participants on various stimuli, thus creating expectations, to test these participants' comprehension against participants not having received training.

8.1 References

We have used Google Scholar to cite references, using the APA style. R and R packages have been cited by using citation() in R.

Arlinger, S., Lunner, T., Lyxell, B., & Kathleen Pichora - Fuller, M. (2009). The emergence of cognitive hearing science. *Scandinavian journal of psychology*,*50*(5), 371-384.

Bates, D., Maechler, M., Bolker, B., & Walker, S. 2015. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-9, URL https://CRAN.R-project.org/package=lme4

Dick, F., Bates, E., & Ferstl, E. C. (2003). Spectral and temporal degradation of speech as a simulation of morphosyntactic deficits in English and German. *Brain and language*, *85*(3), 535-542.

Friedenberg, J., & Silverman, G. (2011). *Cognitive science: An introduction to the study of mind*, Second Edition. Sage, 294-304

Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain research*, *1153*, 134-143.

Herrmann, C. S., Munk, M. H., & Engel, A. K. (2004a). Cognitive functions of gamma-band activity: memory match and utilization. *Trends in cognitive sciences*, *8*(8), 347-355.

Herrmann, C. S., Lenz, D., Junge, S., Busch, N. A., & Maess, B. (2004b). Memory-matches evoke human gamma-responses. *BMC neuroscience*, *5*(1), 13.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal 50(3), 346-363

Hunt, M. J., & Lefebvre, C. (1989, May). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on* (pp. 262-265). IEEE.

Lenz, D., Schadow, J., Thaerig, S., Busch, N. A., & Herrmann, C. S. (2007). What's that sound? Matches with auditory long-term memory induce gamma activity in human EEG. *International Journal of Psychophysiology*, *64*(1), 31-38.

Lutzenberger, W., Pulvermüller, F., & Birbaumer, N. (1994). Words and pseudowords elicit distinct patterns of 30-Hz EEG responses in humans. *Neuroscience letters*, *176*(1), 115-118.

Miettinen, I., Alku, P., Salminen, N., May, P. J., & Tiitinen, H. (2011). Responsiveness of the human auditory cortex to degraded speech sounds: reduction of amplitude resolution vs. additive noise. *Brain research*, *1367*, 298-309.

Miettinen, I., Tiitinen, H., Alku, P., & May, P. J. (2010). Sensitivity of the human auditory cortex to acoustic degradation of speech and non-speech sounds. *BMC neuroscience*, *11*(1), 2-11.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rönnberg, J., Rudner, M., Lunner, T., & Zekveld, A. A. (2010). When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health*, *12*(49), 263.

Schadow, J., Lenz, D., Dettler, N., Fründ, I., & Herrmann, C. S. (2009). Early gamma-band responses reflect anticipatory top-down modulation in the auditory cortex. *Neuroimage*, *47*(2), 651-658.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in neurosciences*, *26*(2), 100-107.

8.2 Footnotes

PsychoPy: http://www.psychopy.org
KorpusDK: http://ordnet.dk/korpusdk_en/facts/available-corpora?set_language=en
Korpus 90: http://ordnet.dk/korpusdk_en/facts/available-corpora/suppliers-of-texts-to-korpus-90?set_language=en
Korpus 2000: http://ordnet.dk/korpusdk_en/facts/available-corpora/suppliers-of-texts-to-korpus-2000?set_language=en
Sample Rate: https://en.wikipedia.org/wiki/Sampling_%28signal_processing%29#Sampling_rate
Apple Logic Pro: http://www.apple.com/dk/logic-pro/
Information(1998-2008) & Wikipedia corpora: http://corp.hum.sdu.dk/cqp.html