# Improving Grapheme-to-Phoneme Conversion for Anglicisms in German Speech Recognition

## Master's Thesis

for the degree MASTER OF SCIENCE in Media Informatics
submitted to the Faculty of Computer Science and Engineering Science
of TH Köln - University of Applied Sciences


**Submitted by** Julia Maria Pritzen (Matr.-No 11092511)
**Address** Goltsteinstraße 89
50968 Köln
julia.pritzen@gmail.com

**First supervisor** Prof. Dr. Dietlind Zühlke
TH Köln
**Second supervisor** Michael Gref, M.Eng.
Fraunhofer IAIS


Technology
Arts Sciences
TH Köln

Cologne, February 12, 2021

# Declaration in lieu of oath

I hereby declare in lieu of an oath that this work is my own and that I have not used any sources other than those listed in the bibliography. Content from published or unpublished works that has been quoted directly or indirectly or paraphrased is indicated as such. The work has not been submitted in the same or similar form or in part for any other academic award. The electronic version I have submitted is completely identical to the hard copy version submitted.

I am aware that my work may be checked for unmarked copying of others' intellectual property for the purpose of a plagiarism check using plagiarism detection software.

I am aware of the punishability of a false Declaration in lieu of an oath, namely the threat of punishment according to § 156 StGB up to three years imprisonment or fine for intentional committal of the offense or according to § 161 Abs. 1 StGB up to one year imprisonment or fine if committed by negligence.

## *Eidesstattliche Erklärung*

*Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.*

*Mir ist bekannt, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs mittels einer Plagiatserkennungssoftware auf ungekennzeichnete Übernahme von fremdem geistigem Eigentum überprüft werden kann.*

*Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.*

Cologne, 02/12/2021

Place, date
*Ort, Datum*

Legally binding signature
*Rechtsverbindliche Unterschrift*

# Abstract

This work designs and evaluates methods for improving the recognition of anglicisms in German speech recognition. Focusing on the pronunciation dictionary of an ASR system, three approaches were designed and implemented for creating supplementary anglicism pronunciation dictionaries. In the first approach, anglicism pronunciations were directly derived from the German Wiktionary. In the second approach, anglicism pronunciations were generated with both a German and an English G2P model. By comparing the confidence measures, the respective best pronunciation was chosen to be added to the resulting anglicism pronunciation dictionary. An additional P2P model was created for this approach that maps English phonemes to their German equivalents. In the third approach, multitask learning was utilized by adding an additional anglicism classification task to a German Seq2Seq G2P model. By distinguishing anglicisms and native German words, the G2P model was able to generate different pronunciations for each respective case. For each resulting anglicism pronunciation dictionary, a dedicated ASR model was created with similar settings. All ASR models including a baseline model were evaluated on a dedicated anglicism test set and two additional German test sets from the broadcast domain to prevent performance issues in other use cases. Ten out of thirteen models performed better than the baseline. The best model resulted from the comparative approach. For the anglicism test set, the WER could be decreased by 0.21 percentage points with 22 more anglicism being recognized compared to the baseline model. The mean WER based on all test sets was decreased by 0.08 percentage points. More anglicism data of better quality and refined model implementations are needed to further improve the anglicism recognition results.

*In dieser Arbeit werden Methoden zur Verbesserung der Erkennung von Anglizismen in der deutschen Spracherkennung konzipiert und evaluiert. Mit Fokus auf dem Aussprachewörterbuch eines ASR-Systems wurden drei Ansätze implementiert, um ergänzende Anglizismen-Aussprachewörterbücher zu erstellen. Im ersten Ansatz wurden die Anglizismen-Aussprachen direkt aus dem deutschen Wiktionary entnommen. Im zweiten Ansatz wurden Anglizismen-Aussprachen sowohl mit einem deutschen als auch mit einem englischen G2P-Modell generiert. Mittels Vergleich der Konfidenzmaße wurde die jeweils beste Aussprache ausgewählt und in das resultierende Anglizismen-Aussprachewörterbuch aufgenommen. Für diesen Ansatz wurde ein zusätzliches P2P-Modell erstellt, das englische Phoneme in ihre deutschen Entsprechungen umwandelt. Im dritten Ansatz wurde Multitask-Learning verwendet, indem einem deutschen Seq2Seq G2P-Modell ein zusätzlicher Task zur Klassifikation von Anglizismen hinzugefügt wurde. Durch die Unterscheidung zwischen Anglizismen und nativen (deutschen) Wörtern konnte das G2P-Modell unterschiedliche Aussprachen für die jeweiligen Fälle generieren. Für jedes resultierende Anglizismen-Aussprachewörterbuch wurde ein eigenes ASR-Modell mit jeweils gleicher Konfiguration erstellt. Alle ASR-Modelle, einschließlich des Baseline-Modells, wurden auf einem dedizierten Anglizismen-Testset sowie zwei zusätzlichen deutschen Testsets aus der Rundfunkdomäne evaluiert, um Probleme in anderen Anwendungsfällen auszuschließen. Zehn von dreizehn Modellen schnitten besser ab als das Baseline-Modell. Das beste Modell resultierte aus dem Vergleichs-Ansatz. Für das Anglizismen-Testset konnte die WER um 0,21 Prozentpunkte gesenkt werden, wobei 22 Anglizismen mehr erkannt wurden als im Baseline-Modell. Die mittlere WER auf Basis aller Testsets wurde um 0,08 Prozentpunkte gesenkt. Um die Ergebnisse der Anglizismenerkennung weiter zu verbessern, werden mehr und hochqualitativere Anglizismen-Daten sowie ausgereiftere Modellimplementierungen benötigt.*

# Contents

**List of Tables** VII

**List of Figures** IX

**List of Listings** X

**List of Samples** XI

**1 Introduction** 1
1.1 The ASR Model at Fraunhofer IAIS . . . 2
1.2 Problem Definition . . . 6
1.3 Relevance . . . 6
1.4 Objective . . . 7
1.5 Limitations . . . 7

**2 Literature Review** 9
2.1 Fundamentals of Grapheme-to-Phoneme Conversion . . . 9
2.1.1 Joint-Sequence G2P Conversion . . . 11
2.1.2 Sequence-to-Sequence G2P Conversion . . . 18
2.2 Wiktionary as Source for Pronunciations . . . 22
2.3 Phoneme Mapping for Non-Native Pronunciation Variants . . . 23
2.4 Multitask Learning . . . 26

**3 Methodology** 31
3.1 Research Questions . . . 31
3.2 Data Collection . . . 31
3.2.1 Anglicism Testset . . . 32
3.2.2 Anglicism List . . . 33
3.2.3 Pronunciation Dictionaries . . . 35
3.2.4 Text-to-Speech-Generated Audio Files . . . 35
3.3 General Evaluation . . . 36
3.3.1 Baseline Model . . . 36
3.3.2 Anglicism Pronunciation Dictionary . . . 36
3.3.3 Benchmarking . . . 37
3.3.4 Evaluation Metrics . . . 39

**4 Approach 1: Using Pronunciations from Wiktionary** 42
4.1 Data Crawling . . . 42
4.2 Dictionary Creation . . . 43
4.3 Evaluation . . . 43
4.3.1 Variation 1: Early Version With Manual Modifications . . . 46
4.3.2 Variation 2: Combination of Both Dictionaries . . . 47
4.3.3 Anglicism Recognition Results . . . 49

**5 Approach 2: Comparing German and English G2P Results** 51
5.1 Creating a P2P model . . . 52

5.1.1 Data Collection . . . 53
5.1.2 Implementation . . . 56
5.1.3 Evaluation . . . 57
5.2 Implementation . . . 61
5.3 Evaluation . . . 63
5.4 Variation: Detecting Anglicisms Based on Crawl Results . . . 67
5.4.1 Data Collection . . . 68
5.4.2 Implementation . . . 69
5.4.3 Evaluation . . . 69
5.5 Anglicism Recognition Results . . . 71

**6 Approach 3: Using Multitask Learning for Anglicism Detection 73**
6.1 Data Collection . . . 73
6.2 Implementation . . . 74
6.3 Tuning . . . 76
6.3.1 Paper Configurations . . . 76
6.3.2 Manual Configurations . . . 76
6.3.3 Hyperparameter Optimization with Optuna . . . 77
6.4 Model Selection . . . 79
6.5 Evaluation . . . 82
6.6 Anglicism Recognition Results . . . 85

**7 Discussion 87**

**8 Conclusion 91**
8.1 Research Questions . . . 92
8.2 Challenges . . . 94

**9 Future Work 98**

**Bibliography 102**

**Appendix 111**

# List of Tables


1.1 Anglicism examples from Fraunhofer IAIS' pronunciation dictionary compared to pronunciations taken from Wiktionary in SAMPA notation. . . . . 7

2.1 Grapheme-to-phoneme alignment for the word "mice" in IPA notation, including one-to-many and null phoneme mapping. . . . . 9
2.2 Results of Sequitur G2P (bottom row) in comparison to other G2P models for the CMUdict dataset. (Bisani and Ney, 2008) . . . . . 18
2.3 Results on the CMUdict dataset including Sequitur G2P baseline. (Yao and Zweig, 2015) . . . . . 21
2.4 Word error rates (%) of ASR systems with *GlobalPhone* & Wiktionary G2P-based dictionaries. (Schlippe et al., 2014) . . . . . 23
2.5 Word error rates (%) using filtered Wiktionary G2P-based dictionaries, highlighting the best result for each language. (Schlippe et al., 2014) . . . . . 23
2.6 PER and WER performance of all tested models for the German PHONOLEX test set. Column "PHONOLEX set" implies which data was used to train the respective model. (Milde et al., 2017) . . . . . 30
2.7 PER and WER performance of selected models specifically for English loanwords inside the German PHONOLEX test set (34 words, 2.59 %). (Milde et al., 2017) . . . . . 30

3.1 Statistics of the test set "Anglicisms 2020". . . . . 34

4.1 WERs for the baseline and wiki-base ASR models. . . . . 44
4.2 Differences in words and pronunciations in the resulting dictionaries. The exclusive words refer to the words that were contained in the respective model's dictionary, but not in the other models dictionary. Similarly, the exclusive pronunciations refer to the pronunciation variations that were contained in the respective model's dictionary, but not in the other models dictionary. . . . 46
4.3 WERs for the baseline, wiki-base and wiki-v1 ASR models. . . . . 47
4.4 WERs for the baseline, wiki-base, wiki-v1 and wiki-v2 ASR models. . . . . 48
4.5 EERs based on a total of 1,362 anglicism entities for the baseline, wiki-base, wiki-v1 and wiki-v2 models after applying the test set "Anglicisms 2020". . . 50

5.1 Best PER results for each the P2P models of each AM weight value after applying the Wiktionary test set. . . . . 58
5.2 PER values of the P2P models of each iteration for AM weight 0.5 after applying the validation set. . . . . 58
5.3 K-fold cross validation results for the models of chosen AM weights. . . . . 59
5.4 PER and WER results of each iteration for the English Sequitur G2P model after applying the test set. . . . . 62
5.5 Pronunciations for the anglicism "Crowdfundings" generated by the comparative ASR models. . . . . 64
5.6 WERs for the baseline and the comparative ASR models. . . . . 64
5.7 Pronunciations for the word "Mirror" in the baseline and comparative dictionaries. . . . . 66

5.8 Typical German words (top 4 rows) and English words (bottom 4 rows) with their confidence measures and pronunciations generated by Fraunhofer IAIS' German G2P model compared to the respective pronunciations taken from Wiktionary, converted to BAS-SAMPA . . . . . 68
5.9 WERs for the baseline and comp-crawl ASR models. . . . . 70
5.10 EERs for the baseline and comparative models after applying the test set "Anglicisms 2020". . . . . 72

6.1 PERs and WERs for the batch size configurations mentioned in Yao and Zweig (2015). . . . . 76
6.2 PERs and WERs for the manual batch size configurations. . . . . 77
6.3 Batch size and learning rate configurations with their resulting PER and WER values of the top five Optuna trial models. . . . . 79
6.4 Selected Seq2Seq MTL models with their data source, number of epochs and number of iterations per epoch. The data source was used to create the train and validation sets. . . . . 80
6.5 Selected MTL models with their G2P task and anglicism classification task evaluation metrics. For models M25-P and O5-P, the precision, recall and F1 score values are 0.00 % because they did not yield any positive classifications. 80
6.6 WERs for the baseline and the MTL ASR models. . . . . 83
6.7 EERs for the baseline and MTL models after applying the test set "Anglicisms 2020". . . . . 86

7.1 All ASR models with their number of anglicism pronunciation dictionary entries, number of total pronunciation dictionary entries (baseline and anglicism pronunciations), number of recognized anglicisms, their WER and EER values for test set "Anglicisms 2020" and a mean WER value that corresponds to the average WER of all test sets ("Anglicisms 2020", "German Broadcast 2020" and "Challenging Broadcast 2018"). . . . . 88
7.2 Five example words with their respective pronunciation(s) from the best performing pronunciation dictionary of each approach. . . . . 89
7.3 WERs of the best models from each approach, including the baseline model for comparison. . . . . 90

8.1 Recognition results for the word "MIDCAP" from CMUdict (Carnegie Mellon University, 2014). The CMUdict pronunciation is the original ARPABET pronunciation that the audio files of the voices were based on. Please note that the pronunciations in rows 2–6 are written in BAS-SAMPA notation. . . . . 96

A1 Videos used in the testset "Anglicisms 2020" . . . . . 113
A2 Examples for the anglicism pronunciation dictionary contents for all Wiktionary models. . . . . 114
A3 WERs per file in test set "Anglicisms 2020" for all Wiktionary models. . . . . 115
A4 WERs per file in test set "German Broadcast 2020" for all Wiktionary models. 116
A5 WERs per file in test set "Challenging Broadcast 2018" for all Wiktionary models. 117
A6 EERs per file in test set "Anglicisms 2020" for all Wiktionary models. . . . . 118
A7 Examples for the anglicism pronunciation dictionary contents of all Comparative models. . . . . 119

A8 WERs per file in test set “Anglicisms 2020” for all Comparative models. . . . . 120
A9 WERs per file in test set “German Broadcast 2020” for all Comparative models. 121
A10 WERs per file in test set “Challenging Broadcast 2018” for all Comparative models. . . . . 122
A11 EERs per file in test set “Anglicisms 2020” for all Comparative models. . . . . 123
A12 Examples for the anglicism pronunciation dictionary contents of all MTL models. 124
A13 WERs per file in test set “Anglicisms 2020” for all MTL models. . . . . 125
A14 WERs per file in test set “German Broadcast 2020” for all MTL models. . . . . 126
A15 WERs per file in test set “Challenging Broadcast 2018” for all MTL models. . 127
A16 EERs per file in test set “Anglicisms 2020” for all MTL models. . . . . 128

# List of Figures

1.1 Overview of the architecture used for the ASR system at Fraunhofer IAIS (Stadtschnitzer, 2018) . . . 3
1.2 Example for a WFST representing a pronunciation dictionary for two different pronunciation variations of the word "tomato" listed in CMUdict (Stadtschnitzer, 2018) . . . 6
1.3 Extract from the Benchmark Viewer at Fraunhofer IAIS showing a manually annotated reference (Ref) and generated hypothesis (Hyp) by the current ASR model for a short audio sample containing the anglicism "Machine Learning". 7

2.1 The grapheme and phoneme sequences for "mixing" split into four (a) and seven (b) graphones. (Bisani and Ney, 2008) . . . 12
2.2 Architecture of an LSTM cell (adapted from Wikimedia Commons, 2015). . . 19
2.3 G2P Encoder-Decoder-LSTM model representation for the grapheme sequence "C A T" with reversed input sequence and phoneme output in ARPABET notation. (Yao and Zweig, 2015) . . . 20
2.4 Transcript FST with included word FSTs (boxes) for the transcript "summer of 69". (adapted from Bruguier et al., 2017) . . . 25
2.5 Single task learning (a) and MTL (b) of four tasks with the same input. (Caruana, 1997) . . . 27
2.6 Hard (a) and soft (b) parameter sharing for multiple tasks. (Ruder, 2017) . 27
2.7 Seq2Seq G2P model with reversed input and additional language marker for the subtask at the start of each word. (Milde et al., 2017) . . . 29

3.1 Manual annotation for the file "Rezo - Die Zerstörung der Presse" in Simple-ELAN. 33
3.2 German inflection box on Wiktionary for the word "downloaden" . . . 34
3.3 Benchmark viewer: Baseline model benchmark result of segment 3 in "code-centric.AI Bootcamp - Was ist Machine Learning" from test set "Anglicisms 2020". . . . 39
3.4 The Benchmark Viewer at Fraunhofer IAIS showing the WERs of the baseline model for different files in test set "Anglicisms 2020". . . . 40

5.1 Process of creating a P2P model that maps English to German pronunciations. Based on an English pronunciation data, German pronunciation equivalents are created by both a G2P and an acoustic model. The best German pronunciation is chosen to be paired with the respective English pronunciation. The pronunciation pair is added to the P2P training data which is used to create the P2P model. . . . 52
5.2 Vowel diagram from the International Phonetic Alphabet (International Phonetic Association, 2015) . . . 60
5.3 Example of the comparative process used on the anglicism "Eyeliner". The grapheme sequence is put into both the German and the English G2P models. The respective output phoneme sequences are then compared to each other by their confidence measure. If the English pronunciation wins, it is mapped to German phonemes by the P2P model. The resulting pronunciation is added to the pronunciation dictionary. . . . 61

6.1 MTL G2P model representation for the grapheme sequence ⟨Fan⟩. The grapheme sequence is processed by the encoder that passes the output to both the decoder and the anglicism classification task. Based on the encoder output, the decoder generates the pronunciation while the classification task asserts the probability for the grapheme sequence being an anglicism. (Yao and Zweig, 2015, adapted from) . . . . . 75

7.1 Correlation of WER and EER values of test set “Anglicisms 2020” for all ASR models. . . . . 89

A1 Confusion matrices showing the relative anglicism classification results of all MTL models after applying the respective test set. The y-axis shows the actual classification while the x-axis shows the classification predicted by the model. “Yes” and “No” states if the grapheme sequence was classified as an anglicism or not. . . . . 111

A2 Confusion matrices showing the absolute anglicism classification results of all MTL models after applying the respective test set. The y-axis shows the actual classification while the x-axis shows the classification predicted by the model. “Yes” and “No” states if the grapheme sequence was classified as an anglicism or not. . . . . 112

# List of Listings

1.1 Example for multiple pronunciation variations and homophones in CMUdict (Carnegie Mellon University, 2014) . . . . . 4

2.1 Discounted EM algorithm for the training process of the Sequitur G2P model in pseudocode. (Bisani and Ney, 2008; edited to match equation order) . . . 17

3.1 Entity notation in the annotation of “Rezo - Die Zerstörung der Presse” that marks the anglicisms “Social Media” and “Storys”. . . . . 33
3.2 Formatting example of the baseline pronunciation dictionary. . . . . 36
3.3 JSON format: Baseline model benchmark result of segment 3 in “Was ist Machine Learning, eine Einführung - codecentric.AI Bootcamp” from test set “Anglicisms 2020”. . . . . 38

4.1 Extract from the Wiktionary anglicism dictionary . . . . . 43

5.1 Original CMUdict pronunciations and cleaned versions without stress markers and variant brackets. . . . . 54
5.2 Different German Sequitur G2P outputs for words in uppercase and title case. 55
5.3 JSON output by the live recognizer. . . . . 55
5.4 Bash command to train the first iteration of a P2P model with Sequitur G2P. . 56
5.5 Selection of the final German pronunciation equivalent in pseudocode. . . . . 57

# List of Samples


3.1 Segment 3 in “codecentric.AI Bootcamp - Was ist Machine Learning” from test set “Anglicisms 2020”. . . . . 40

4.1 Segment 15 in “Rezo - Die Zerstörung der CDU” from test set “Anglicisms 2020”. 44
4.2 Segment 0 in “Reportage vom Inneren der Bgm.-Smidt-Brücke” from test set “Challenging Broadcast 2018”. . . . . 45
4.3 Segment 7 in “Rezo - Wie Politiker momentan auf Schüler scheißen” from test set “Anglicisms 2020”. . . . . 45
4.4 Segment 13 in “Rezo - Die Zerstörung der CDU” from test set “Anglicisms 2020”. 47
4.5 Segment 39 in “tagesschau 20:00 Uhr, 03.02.2020” from test set “Anglicisms 2020”. . . . . 48
4.6 Segment 25 in “tagesthemen 22:15 Uhr, 18.02.2020” from test set “Anglicisms 2020”. . . . . 49
4.7 Segment 13 in “Venix - TechNews 94” from test set “Anglicisms 2020”. . . . . 49

5.1 Segment 9 in “Venix - TechNews 94” from test set “Anglicisms 2020”. . . . . 65
5.2 Segment 19 in “Rezo - Wie Politiker momentan auf Schüler scheißen” from test set “Anglicisms 2020”. . . . . 66
5.3 Segment 16 in “Der Limbecker Platz in Essen” from test set “German Broadcast 2020”. . . . . 67
5.4 Segment 8 in “Polizeigewalt gegen Demonstranten in Hongkong” from test set “German Broadcast 2020”. . . . . 70
5.5 Segment 39 in “Venix - Tech News 95” from test set “Anglicisms 2020”. . . . 71
5.6 Segment 24 in “Die Heldenreise des Pre-Sales Consultant” from test set “Anglicisms 2020”. . . . . 71

6.1 Segment 1 in “Wirtschaft regional - Teeherstellung in Bremen” from test set “Challenging Broadcast 2018”. . . . . 83
6.2 Segment 13 in “Besuch von Bundeskanzlerin Merkel in Indien” from test set “German Broadcast 2020”. . . . . 84
6.3 Segment 57 in “Rezo - Die Zerstörung der Presse” from test set “Anglicisms 2020”. . . . . 84
6.4 Segment 51 in “Rezo - Die Zerstörung der CDU” from test set “Anglicisms 2020”. 85

9.1 Segment 37 in “tagesthemen 22:15 Uhr, 18.02.2020” from test set “Anglicisms 2020”. . . . . 100

Chapter 1

# Introduction

**Contents**


1.1 The ASR Model at Fraunhofer IAIS . . . . . 2
1.2 Problem Definition . . . . . 6
1.3 Relevance . . . . . 6
1.4 Objective . . . . . 7
1.5 Limitations . . . . . 7


Automatic speech recognition (ASR) is the process of automatically transcribing speech to text. With the increasing popularity of Amazon's Alexa, Apple's Siri and Google's Assistant, ASR has become more accessible to the public. The transformation of spoken to written language offers advantages in many domains. While ASR is used for the aforementioned voice assistant systems or to increase accessibility for hearing impaired people, it can also be used to transcribe audio recordings for making the spoken content searchable in large media archives.

Among other services in the segment of speech technologies, the Fraunhofer Institute for Intelligent Analysis and Information Systems (Fraunhofer IAIS) developed an ASR system that is specialized on the German broadcast domain (e.g. Schmidt, 2020), parliamentary sessions (e.g. Klatte, 2020) and oral history interviews (Gref et al., 2018). Primarily trained on audio data from German TV and radio segments, it offers state-of-the-art speech recognition for numerous clients. However, the ASR system came across an issue that has yet to be solved: the reliable recognition of anglicisms in German speech.

Görlach (1994) defines an anglicism as follows:

> An Anglicism is a word or idiom that is recognizably English in its form (spelling, pronunciation, morphology, or at least one of the three), but is accepted as an item in the vocabulary of the receptor language.
>
> **(Görlach, 1994, p.224)**

Anglicisms are becoming increasingly common in the German language. From 1994 to 2004, the use of anglicisms has doubled (Burmasova, 2010). In 2013, anglicisms

accounted for 3.5 % of all words in the German dictionary Duden (RP Online, 2013). In a recent study by Hunt (2019), Anglicisms made up 4.53 % of all word types based on a corpus of everyday spontaneous German speech samples. In contrast to native German words, anglicisms are often pronounced differently due to their English heritage. This proposes a challenge for a monolingual German ASR model as the inventory of recognizable pronunciations is mainly based on German pronunciation rules.

## 1.1 The ASR Model at Fraunhofer IAIS

An automatic speech recognition system generates a sequence of textual recognition hypotheses from a speech audio segment. The ASR model used at Fraunhofer IAIS is a Hidden-Markov-Model (HMM) based recognizer which is considered a statistical speech recognition system. (Stadtschnitzer, 2018)

A statistical speech recognition system is based on the Bayes' theorem (Bayes, 1763; Stadtschnitzer, 2018):

$$p(w_1^N|x_1^T) = \frac{p(x_1^T|w_1^N) \cdot p(w_1^N)}{p(x_1^T)} \tag{1.1}$$

Given a sequence[1] $x_1^T = (x_1, ..., x_t)$ of acoustic features, it tries to find the sequence of words $w_1^N = (x_1, ..., x_n)$ that maximizes the posterior probability over $w_1, ..., w_N$ (Ney and Ortmanns, 1999). Figure 1.1 on the following page shows the architecture of the ASR system used at Fraunhofer IAIS.

The ASR system consists of the following components:

### Feature Extraction

Discriminative features $x_1^T$ are extracted out of the raw speech input that can provide helpful information for the model's learning process. An example for prominent features are *Mel-Frequency Cepstral Coefficients* or *Filterbank Coefficients* which represent the speech signal's spectrogram[2]. (Stadtschnitzer, 2018)

### Acoustic Model

The acoustic model provides "stochastic models that capture both the temporal and static features of the speech signal" (Stadtschnitzer, 2018, p.25). Instead of modeling whole

[1]An alternative notation for sequence indexing which has also been applied in Ney and Ortmanns (1999) is used to make the equations more readable. $x_1^T$ is the alternative notation for $(x_t)_{t=1}^T$; $w_1^N$ is the alternative notation for $(w_n)_{n=1}^N$.

[2]The spectrogram shows a visual representation of an audio signal's frequency spectrum by time

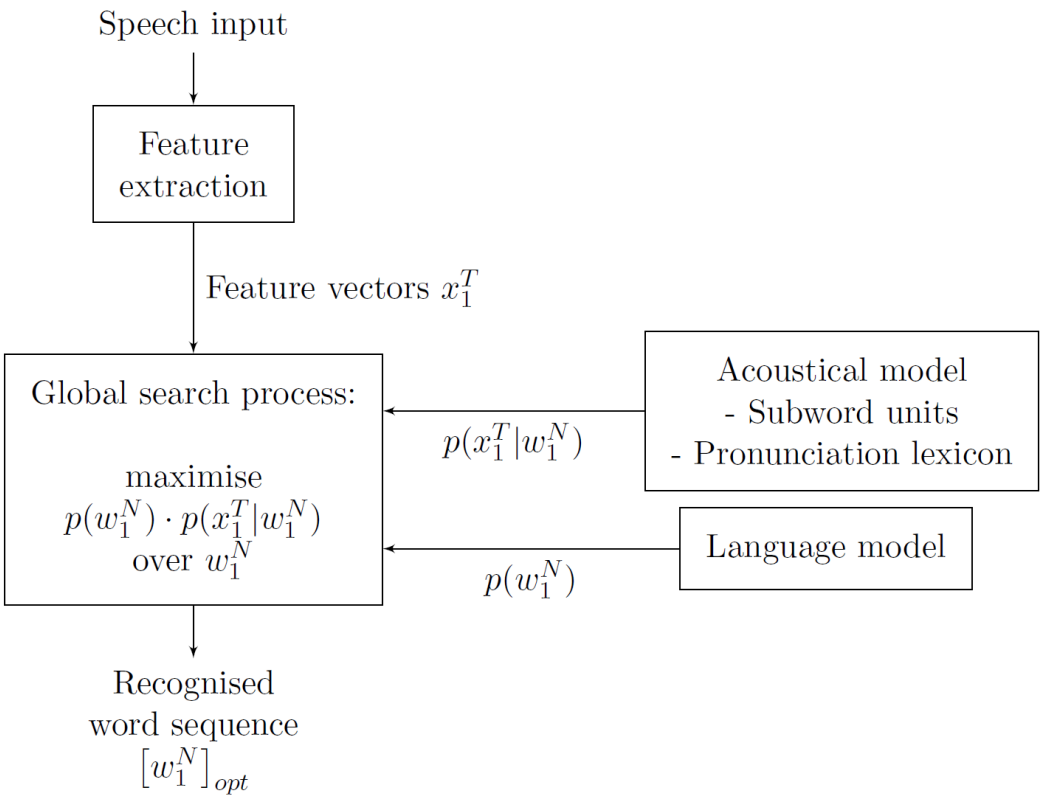


**Figure 1.1:** Overview of the architecture used for the ASR system at Fraunhofer IAIS (Stadtschnitzer, 2018)

words, phonemes[1] are used as subword units for a better generalization. By doing this, the model not only has more training data available, but it is also more versatile because "the vocabulary can be gracefully extended independent of the acoustic training data" (Stadtschnitzer, 2018, p.23) since the words within a language all use the same set of phonemes. Given the speech input, the acoustic model concatenates phoneme models using HMMs. The HMMs make it able to handle variations such as a varying speaking rate in the speech signal. The acoustic model is trained by speech waveforms and their orthographic transcriptions. The output of the acoustic model $p(x_1^T|w_1^N)$ is the "probability of observing the feature sequence $x_1^T$ given the word sequence $w_1^N$" (Stadtschnitzer, 2018, p.25). (Stadtschnitzer, 2018; Yao and Zweig, 2015)

At Fraunhofer IAIS, a hybrid approach is applied for the acoustic model where deep neural networks are used to estimate observation probabilities of HMMs (Yu and Deng, 2014).

[1] A phoneme is considered the "smallest unit of sound that distinguishes one word from another in a particular language" (Stadtschnitzer, 2018, p.23). In the field of speech science, however, the term *phoneme* is often misused with the term *phone* (Moore and Skidmore, 2019). While phonemes are a "mental representation in the mind of a speaker"(Reetz and Jongman, 2020, p.26), phones are defined as the actual physical representations of speech sounds. Since this distinction is not made in most literature used, the term phoneme will be used for both definitions in this work. More on this topic can be read in Moore and Skidmore (2019).

More specifically, LF[1]-MMI[2] models (Povey et al., 2016) are applied using TDDN[3]-LSTM topology. (Gref et al., 2020)

### Pronunciation Dictionary

The pronunciation dictionary is a part of the acoustic model. It contains mappings between words and their respective pronunciation. Each pronunciation is separated into phonemes. The phonemes are usually written in a machine-readable phonetic alphabet notation like SAMPA or ARPAbet. A word can be listed multiple times with different pronunciation variations. Also, a pronunciation can be mapped to different words in the case of homophones. Listing 1.1 shows an example for multiple pronunciation variations and homophones in the English pronunciation dictionary CMUdict. (Stadtschnitzer, 2018; Yao and Zweig, 2015)

Building a pronunciation dictionary can be done manually by professional linguists, but considering the potentially large vocabulary of a language, it is a very time-consuming and cost-intensive task. However, this task can be automated using machine learning methods. Grapheme-to-Phoneme (G2P) models are able to generate pronunciations (phoneme sequences) from words (grapheme[4] sequences). Trained on an existing pronunciation dictionary like PHONOLEX or CMUdict, a G2P model finds grapheme-phoneme-pairs (*graphones* (Bisani and Ney, 2008)) for translating seen or unseen words to their respective pronunciations. With the help of this data-driven method, small pronunciation dictionaries

**Listing 1.1:** Example for multiple pronunciation variations and homophones in CMUdict (Carnegie Mellon University, 2014)

```
# word with multiple pronunciation variations
EVOLUTION  EH2 V AH0 L UW1 SH AH0 N
EVOLUTION  IY2 V AH0 L UW1 SH AH0 N
EVOLUTION  EH2 V OW0 L UW1 SH AH0 N
EVOLUTION  IY2 V OW0 L UW1 SH AH0 N
...
# homophones
PAUSE  P AO1 Z
...
PAWS  P AO1 Z
```

[1] **Lattice-Free:** Without the use of word lattices which represent an approximation for all possible word sequences in the language model

[2] **Maximum Mutual Information:** Maximizing "the conditional [globally normalized] log-likelihood of the correct transcript" (Povey et al., 2016)

[3] **Time Delay Neural Network:** A network able to process temporal dependencies by using time frames on multiple inputs

[4] A grapheme is the smallest unit of writing that corresponds to a phoneme

can be automatically extended by applying the G2P model to large collections of words. (Stadtschnitzer, 2018; Bisani and Ney, 2008)

### Language Model

The language model is an $n$-gram model that determines the probability of each word on its $n-1$ predecessors. It "models the probabilities of sentences (including the semantics and the syntax) of the considered language" (Stadtschnitzer, 2018, p.22). The language model is trained on large amounts of example texts by counting $n$-gram occurrences to form maximum likelihood parameter estimates. The higher the value for $n$, the less likely it is that all possible $n$-grams are found during training, resulting in a probability of 0 for unseen $n$-grams. To avoid this problem when having insufficient data, smoothing algorithms like *Kneser-Ney smoothing* (Kneser and Ney, 1995) can be used to modify the $n$-gram probability distribution. The output of the language model $p(w_1^N)$ is the total probability of a word sequence $w_1^N$. (Stadtschnitzer, 2018; Yao and Zweig, 2015)

### Global Search Process

Given the output probabilities of the acoustic model and language model, the search process finds the word sequence $[w_1^N]_{opt}$ that is most probable according to the following equation:

$$\begin{aligned}[w_1^N]_{opt} &= \underset{w_1^N}{\operatorname{argmax}}\{p(w_1^N|x_1^T)\} \\ &= \underset{w_1^N}{\operatorname{argmax}}\{p(x_1^T|w_1^N)\cdot p(w_1^N)\}\end{aligned} \tag{1.2}$$

where $w_1^N$ is a word sequence, $p(x_1^T|w_1^N)$ is the probability for the word sequence based on the feature sequence $x_1^T$ according to the acoustic model and $p(w_1^N)$ is the probability for the word sequence according to the language model. As this equation originates from Equation (1.1) on page 2, it seems like $p(x_1^T)$ from the numerator is missing. Since $p(x_1^T)$ is a constant value which is not dependent on $w$, it can be disregarded in this case. (Stadtschnitzer, 2018)

To find the most probable sequence among the list of hypotheses, several techniques are applied. Since many of the hypotheses have common subsequences, the *Viterbi decoding algorithm* (Viterbi, 1967) is used to find the best acoustic model probabilities. Also, pruning techniques are used to lighten resource-intensive ASR tasks. (Stadtschnitzer, 2018)

Since Fraunhofer IAIS' ASR system used the Kaldi toolkit, weighted finite state transducers

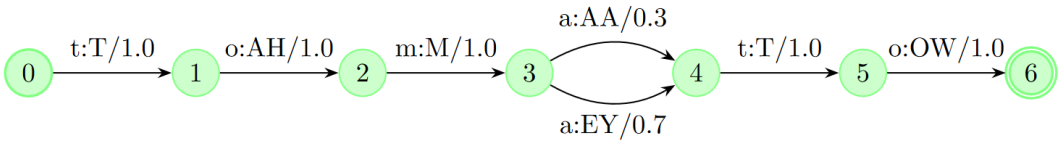


**Figure 1.2:** Example for a WFST representing a pronunciation dictionary for two different pronunciation variations of the word "tomato" listed in CMUdict (Stadtschnitzer, 2018)

(WFSTs) are used. An FST is a "finite automaton whose state transitions are labelled [sic] with both input and output symbols" (Stadtschnitzer, 2018, p.29). A path through the transducer corresponds to a mapping between an input and an output sequence. WFSTs additionally assign weights on the transitions to consider measures like probabilities, penalties and durations. In Kaldi, the HMMs of the acoustic model, the pronunciation dictionary as well as the language model are implemented as WFSTs. Figure 1.2 shows an example of a WFST representing a pronunciation dictionary. Finally, those sub-WFSTs are integrated into a summarized WFST for the whole ASR decoding algorithm to find the most probable word sequence $[w_1^N]_{opt}$. (Stadtschnitzer, 2018)

## 1.2 Problem Definition

The set of phonemes a G2P model uses for the pronunciation generation depends on the training data. Typically, the training data consists of a single language. A model trained with German data will therefore consist of phonemes used in the German language. If this model is used on an English word, the rules learned by the German training data will be applied. This often results in the generation of wrong pronunciations for anglicisms which leads to wrong entries in the pronunciation dictionary. Wrong entries in the pronunciation dictionary ultimately lead to errors in the ASR recognition results. Hence, a solution is needed to improve the generation of anglicism pronunciations.

## 1.3 Relevance

Anglicisms are appearing more frequently in the German language year after year (Burmasova, 2010). Recently, Duden caused a small controversy because anglicisms accounted for a large portion of the 3,000 newly added words in their 28th edition (Jedicke, 2020). Anglicisms have become a relevant part of the German language, but the challenge of understanding them in ASR is a problem that does not yet have a published standard solution.

In their own ASR systems, Fraunhofer IAIS frequently experiences challenges with anglicisms. Table 1.1 on the following page shows examples of anglicisms taken from their

| Word | IAIS Pronunciation | Wiktionary Pronunciation |
|---|---|---|
| Coffee | k O f e: | k O f i |
| geleakt | g @ l Q a k t | g @ l i: k t |
| jammen | j a m @ n | dZ E m @ n |
| Mixtape | m I k s t a: p @ | m I k s t e: p |
| nice | n I s | n aI s |
| rappen | r a p @ n | r E p n |
| Whistleblower | v I s t l e: p l o 6 | v I s l b l O U6 |

**Table 1.1:** Anglicism examples from Fraunhofer IAIS' pronunciation dictionary compared to pronunciations taken from Wiktionary in SAMPA notation.

| | |
|---|---|
| Ref | es geht beim Machine Learning nicht nur um Mathematik . |
| Hyp | es geht beim Maschinenbauer einigen nicht nur um Mathematik |

**Figure 1.3:** Extract from the Benchmark Viewer at Fraunhofer IAIS showing a manually annotated reference (Ref) and generated hypothesis (Hyp) by the current ASR model for a short audio sample containing the anglicism "Machine Learning".

current pronunciation dictionary[1]. Incorrect or missing pronunciations in the dictionary can cause faulty text hypotheses by the ASR system (see Figure 1.3).

## 1.4 Objective

In this work, the recognition of anglicisms in German ASR will be investigated. Methods of improving anglicism recognition will be designed, implemented and evaluated. Several experiments will be conducted that each result in an anglicism pronunciation dictionary. By adding the dictionary to an existing ASR model, the impact of the added pronunciations on anglicism recognition will be measured. Depending on the success of the experiments, the best dictionary will be added to Fraunhofer IAIS' own ASR system to improve the recognition of anglicisms in German ASR.

## 1.5 Limitations

This work focuses on improving the pronunciation dictionary. The other components of the ASR model remain unaffected. Although only the detection of anglicisms in German

[1]As a reference, a short overview of the German pronunciation including SAMPA symbols and respective audio examples can be found at `http://www.coli.uni-saarland.de/elaut/Languages_Sites/sampaDeutsch.htm`

ASR is examined, the results should be applicable to loanwords of other languages as well. Those, however, will not be evaluated in this work. The code created for this work cannot be published as it contains proprietary data and components from Fraunhofer IAIS. However, all steps necessary to reproduce the experiments are described.

# Chapter 2

# Literature Review

**Contents**


2.1 Fundamentals of Grapheme-to-Phoneme Conversion . . . . . 9
2.2 Wiktionary as Source for Pronunciations . . . . . 22
2.3 Phoneme Mapping for Non-Native Pronunciation Variants . . . . . 23
2.4 Multitask Learning . . . . . 26


This chapter describes the current state of research and reviews relevant and related literature. The sections present methods that help answering the research questions and provide a better understanding for the techniques used in this work.

## 2.1 Fundamentals of Grapheme-to-Phoneme Conversion

When reading out loud an unknown word, humans instinctively use their native languages rules to identify the unknown word's pronunciation. Based on the letters and their combination, the corresponding pronunciation is determined. This behavior can be applied to machines as well using Grapheme-to-Phoneme (G2P) conversion.

G2P conversion is a concept where a sequence of letters (graphemes) is translated to a pronunciation represented by a sequence of phonemic transcriptions (phonemes). One phoneme unit can be represented as many grapheme units (e.g. /ʃ/ → ⟨sh⟩) and vice versa (e.g. /eɪ/ → ⟨a⟩). Also, a grapheme might not correspond to a phoneme at all, which is called a *null phoneme*. Table 2.1 shows a phoneme alignment example for the word "mice". (Yao and Zweig, 2015)

Pronunciation dictionaries that have been manually reviewed by professional linguists (e.g. PHONOLEX core) are usually used for training a G2P model. Typically, the training data is formatted having one word with its respective pronunciation per line, separated by a delimiter. The phonemes in the pronunciation are split by whitespaces because the notations SAMPA and ARPABET, both ANSII-expressions of the International Phonetic

| Graphemes | M | I | C | E |
|---|---|---|---|---|
| Phonemes | m | aɪ | s | *null* |

**Table 2.1:** Grapheme-to-phoneme alignment for the word "mice" in IPA notation, including one-to-many and null phoneme mapping.

Alphabet (IPA) for machine-readability, may consist of multiple characters per phoneme. Based on the training data, the model learns how to automatically align input graphemes to their most probable output phonemes. The final model can be used to generate pronunciations for words that do not have a manual transcription, saving time and resources by automatically handling large lists of words and thereby making it possible to significantly extend a pronunciation dictionary for an ASR system.

To evaluate a G2P model, the phoneme error rate (PER) and word error rate (WER) are calculated. The PER states how many predicted single phonemes within a result sequence did not match the expected ones. The WER states how many of the whole predicted phoneme sequences did not match the expected ones. For example, when comparing the predicted phoneme sequence $/\text{maʊz}/$ to the expected sequence $/\text{maɪs}/$, the prediction contains two phoneme errors ($/\text{ʊ}/$ & $/\text{z}/$) and one word error as the whole sequence does not match the expectation. Usually, the PER is lower than the WER since just one wrong phoneme in a sequence of 15 total phonemes leads to a word error, but it only has one phoneme error among 14 correct ones.

There are several techniques for G2P conversion. The most simple approach is to manually create pronunciations for the corresponding words. The skill of linguist professionals is needed for this task to guarantee high quality results. Since a dictionary of significant size is needed, this would be a tedious and costly task. A more automated technique is the *rule-based* approach where rules are defined to match grapheme sequences to corresponding phonemes. The drawback of this approach is that designing the rules is hard and has to be done by professional linguists. Also, since most languages show some irregularities, exceptions in form of a lexicon or special rules have to be created to catch irregularities. (Bisani and Ney, 2008)

Because of the complexity of a language, even linguist professionals might not be able to agree on a uniform linguistic model which will result in a slightly subjective view of the language. A more objective technique that does not rely on knowledge like the two previous approaches is the *data-driven* approach where G2P alignments are predicted by analogy. Data-driven G2P alignment can be done by local classification (e.g. Häkkinen et al. 2003), nearest-neighbor-like approaches (e.g. Bellegarda 2005), regression trees (e.g. Jiang et al. 1997) and other probabilistic approaches. (Bisani and Ney, 2008)

Two different G2P approaches will be utilized in this work:

- **Joint-Sequence G2P conversion**
  The first joint-sequence G2P model that has been developed was introduced by Bisani and Ney (2008). This model is still commonly used in the ASR field (Milde et al., 2017) and often used as a baseline when introducing new G2P approaches,

e.g. by Yao and Zweig (2015). As it is also used at Fraunhofer IAIS, Sequitur G2P has been chosen to be used in this work.

- **Sequence-to-Sequence G2P conversion**
  Sequence-to-Sequence (Seq2Seq) G2P conversion is a Deep Neural Network (DNN) approach that has first been introduced by Yao and Zweig (2015) for the G2P task. Built as an encoder-decoder architecture, it utilizes Long-Short-Term-Memory (LSTM) cells to map grapheme sequences to phoneme sequences. This model has been chosen to be used in this work to cover a DNN approach that enables additional possibilities like Multitask Learning (see Section 2.4 on page 26) that may help with anglicism detection.

The following subsections describe those two techniques in more detail.

### 2.1.1 Joint-Sequence G2P Conversion

Joint sequence models are a probabilistic framework for finding the pronunciation (phoneme sequence) of a given word (grapheme sequence). As joint sequence models have first been introduced by Bisani and Ney (2008), their Sequitur G2P model will be described in this subsection. The input sequence consists of graphemes from a set of graphemes $G$ and the output sequence consists of phonemes from a set of phonemes $\phi$. Using Bayes' decision rule, Bisani and Ney formalize the G2P conversion task as

$$\varphi(g) = \underset{\varphi' \in \phi^*}{\operatorname{argmax}}\, p(g, \varphi'). \tag{2.1}$$

In this equation, the most likely pronunciation $\varphi \in \phi^*$ is found for a grapheme sequence $g \in G^*$. The *Kleene star* operation is used on both the grapheme ($G^*$) and phoneme set ($\phi^*$) to notate the set of all strings over symbols in $G$ and $\phi$ respectively, including the empty string $\varepsilon$. This decision strategy minimizes the risk of not getting the correct pronunciation with respect to word errors. (Bisani and Ney, 2008)

The idea of a joint-sequence model is that "the relation of input and output sequences can be generated from a common sequence of joint units which carry both input and output symbols" (Bisani and Ney, 2008). The simplest case is the one of an FST where each unit contains zero or one input as well as output symbols. In the case of Bisani and Neys G2P model, each unit contains multiple input and output symbols. A unit is then referred to as a *joint multigram* or *graphone*. A graphone $q$ is defined as a pair of a grapheme and a phoneme sequence of potentially different lenghts:

$$q = (g, \varphi) \in Q \subseteq G^* \times \phi^* \tag{2.2}$$

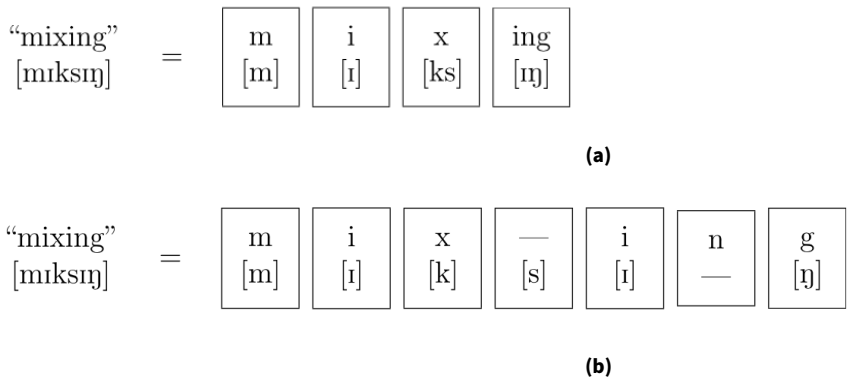


**Figure 2.1:** The grapheme and phoneme sequences for "mixing" split into four (a) and seven (b) graphones. (Bisani and Ney, 2008)

The set of graphones $Q$ can either be defined manually or it can be automatically obtained from the training set. For each word, it is assumed that the grapheme and phoneme sequence are generated by a common graphone sequence. Figure 2.1a shows the word "mixing" as a sequence of four graphones. Here, the grapheme and phoneme sequences are grouped into an equal number of segments. (Bisani and Ney, 2008)

The grouping itself is called *co-segmentation*. An equally valid co-segmentation is shown in Figure 2.1b, which is called *FST-type aligmnent* since each unit contains zero or one input and output symbol. Generally, the alignment in a joint sequence model is referred to as ambiguous $m$-to-$n$ alignment because the input sequence can be freely grouped. Therefore, the joint probability is calculated as

$$p(\boldsymbol{g}, \boldsymbol{\varphi}) = \sum_{\boldsymbol{q} \in S(\boldsymbol{g}, \boldsymbol{\varphi})} p(\boldsymbol{q}), \tag{2.3}$$

where $\boldsymbol{q} \in Q^*$ is a sequence[1] of graphones and $S(\boldsymbol{g}, \boldsymbol{\varphi})$ is the set of all $n$-to-$m$ alignments of $\boldsymbol{g}$ and $\boldsymbol{\varphi}$:

$$S(\boldsymbol{g}, \boldsymbol{\varphi}) := \left\{ \boldsymbol{q} \in Q^* \,\middle|\, \begin{array}{l} \boldsymbol{g}_{q_1} \smile \ldots \smile \boldsymbol{g}_{q_K} = \boldsymbol{g} \\ \boldsymbol{\varphi}_{q_1} \smile \ldots \smile \boldsymbol{\varphi}_{q_K} = \boldsymbol{\varphi} \end{array} \right\} \tag{2.4}$$

where $\smile$ stands for the sequence concatenation and $K = |\boldsymbol{q}|$ is the length of the graphone sequence $\boldsymbol{q}$. The joint probability distribution $p(\boldsymbol{g}, \boldsymbol{\varphi})$ has been reduced to the graphone probability distribution $p(\boldsymbol{q})$ over the sequences $\boldsymbol{q} = (q_1, \ldots, q_K)$ accordingly. The following equation shows how it is modeled using M-gram approximation:

[1] Bisani and Ney (2008) use a bold notation to represent sequences. Since their work is referenced, this notation will be applied in all equations within this subsection.

$$p(\boldsymbol{q}) \cong \prod_{j=1}^{K+1} p(q_j | q_{j-1}, ..., q_{j-M+1}) \tag{2.5}$$

Positions $i < 1$ and $i > K$ are filled with a special boundary symbol $q_i = \perp$ to mark the start and end of a sequence. This allows the modeling of special phenomena at word starts and ends, e.g. terminal devoicing in the German language. (Bisani and Ney, 2008)

The model estimation consists of several steps which will be explained in the following paragraphs. The training data consists of $N$ words and their respective pronunciations:

$$\mathcal{O}_1, ..., \mathcal{O}_N = (\boldsymbol{g}_1, \boldsymbol{\varphi}_1), ..., (\boldsymbol{g}_N, \boldsymbol{\varphi}_N) \tag{2.6}$$

where $\mathcal{O}_k$ is a single training sample. Right now, the training data only consists of raw grapheme and phoneme sequences which have yet to be aligned on the on the level of letters and phonemes. In a joint sequence model, the probability of any co-segmentation can be calculated for each sample. Every joint sequence is uniquely defined by a co-segmentation $\mathcal{S}$ which is used as a a hidden variable:

$$p(\boldsymbol{g}, \boldsymbol{\varphi}, \mathcal{S}) = p(\boldsymbol{q}) \tag{2.7}$$

In this equation, the segmentation into joint units $\mathcal{S}$ has been added to the joint probability distribution $p(\boldsymbol{g}, \boldsymbol{\varphi})$ to define the respective co-segmentation. The log likelihood of the training data is defined as the sum over all segmentations:

$$\begin{aligned} log\mathcal{L}(\mathcal{O}_1, ..., \mathcal{O}_N) &= \sum_{i=1}^{N} log\mathcal{L}(\mathcal{O}_1) \\ &= \sum_{i=1}^{N} log\left( \sum_{\mathcal{S} \in S(\mathcal{O}_i)} p(\mathcal{O}_i, \mathcal{S}) \right) \end{aligned} \tag{2.8}$$

where $\mathcal{L}$ is the likelihood, $\mathcal{O}$ is the respective set of training data and the co-segmentation $\mathcal{S}$ is applied as a hidden parameter for the co-segmentation. (Bisani and Ney, 2008)

Using the expectation maximization (EM) algorithm, maximum likelihood training of the model is performed. First, the context independent unigram case ($M = 1$) is considered. The re-estimation equations are changed accordingly. In Equation (2.9) on the following page, the parameter set $\boldsymbol{\vartheta}$ is added to the graphone probability distribution:

$$p(\boldsymbol{q};\boldsymbol{\vartheta}) = \prod_{j=1}^{|\boldsymbol{q}|} p(q_j|\boldsymbol{\vartheta}) \tag{2.9}$$

The evidence for $q$ is the expected number of occurrences of the graphone $q$ in the training sample under the current set of parameters $\boldsymbol{\vartheta}$. It is calculated as follows:

$$\begin{aligned} e(q;\boldsymbol{\vartheta}) &:= \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i;\boldsymbol{\varphi}_i)} p(\boldsymbol{q}|\boldsymbol{g}_i,\boldsymbol{\varphi}_i;\boldsymbol{\vartheta}) n_q(\boldsymbol{q}) \\ &= \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i;\boldsymbol{\varphi}_i)} \frac{p(\boldsymbol{q};\boldsymbol{\vartheta})}{\sum_{\boldsymbol{q}'\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} p(\boldsymbol{q}';\boldsymbol{\vartheta})} n_q(\boldsymbol{q}) \end{aligned} \tag{2.10}$$

where $e(q;\boldsymbol{\vartheta})$ is the the evidence for $q$, $\boldsymbol{\vartheta}$ is the current parameter set and $n_q(\boldsymbol{q})$ is the number of occurrences of graphone $q$ in sequence $\boldsymbol{q}$. The evidence is calculated by a forward–backward procedure after Deligne and Bimbot (1997). (Bisani and Ney, 2008)

The graphone probability distribution for the updated parameter set $\boldsymbol{\vartheta}'$ is defined as

$$p(q;\boldsymbol{\vartheta}') = \frac{e(q;\boldsymbol{\vartheta})}{\sum_{q'} e(q';\boldsymbol{\vartheta})} \tag{2.11}$$

For models of higher order ($M > 1$), the variable $\boldsymbol{h}$ is used to denote the sequence of preceding joint units $h_j = (q_{j-M+1}, ..., q_{j-1})$. $n_{q,\boldsymbol{h}}(\boldsymbol{q})$ is defined to assign the number of $M$-gram occurrences $q_{j-M+1}, ..., q_j$ in $\boldsymbol{q}$. Variable $\boldsymbol{h}$ is hence added to the re-estimation equations. The graphone probability distribution including the sequence of preceding joint units is defined as follows: (Bisani and Ney, 2008)

$$p(\boldsymbol{q};\boldsymbol{\vartheta}) = \prod_{j=1}^{|\boldsymbol{q}|} p(q_j|h_j;\boldsymbol{\vartheta}) \tag{2.12}$$

When calculating the the evidence for $q$, the number of occurrences of graphone $q$ in sequence $\boldsymbol{q}$ now includes the $\boldsymbol{h}$ variable:

$$\begin{aligned} e(q,\boldsymbol{h};\boldsymbol{\vartheta}) &:= \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i;\boldsymbol{\varphi}_i)} p(\boldsymbol{q}|\boldsymbol{g}_i,\boldsymbol{\varphi}_i;\boldsymbol{\vartheta}) n_{q,\boldsymbol{h}}(\boldsymbol{q}) \\ &= \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i;\boldsymbol{\varphi}_i)} \frac{p(\boldsymbol{q};\boldsymbol{\vartheta})}{\sum_{\boldsymbol{q}'\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} p(\boldsymbol{q}';\boldsymbol{\vartheta})} n_{q,\boldsymbol{h}}(\boldsymbol{q}) \end{aligned} \tag{2.13}$$

The graphone probability distribution including $\boldsymbol{h}$ for the updated parameter set $\boldsymbol{\vartheta}'$ is hence defined as

$$p(q|\boldsymbol{h};\boldsymbol{\vartheta}') = \frac{e(q,\boldsymbol{h};\boldsymbol{\vartheta})}{\sum_{q'} e(q',\boldsymbol{h};\boldsymbol{\vartheta})} \tag{2.14}$$

Once the probability of a new graphone is zero, it cannot emerge in the model. Because of that, the model parameters are initialized by "assigning a uniform distribution over all graphones satisfying certain manually set length constraints" (Bisani and Ney, 2008). Typically, a simpler upper limit $L$ is used (i.e. $|\boldsymbol{g}_p| \leq L$ and $|\boldsymbol{\varphi}_p| \leq L$) and the case $|\boldsymbol{g}_p| = |\boldsymbol{\varphi}_q| = 0$ is excluded. More complex constraints are possible as well, e.g. an additional lower limit. Generally, the initial distribution is defined by the inverse of the total number of allowed graphones:

$$p_0(q) = \left[\sum_{l=0}^{L}\sum_{r=0}^{L}|G|^l|\phi|^r\right]^{-1} \tag{2.15}$$

where the summand for $r = l = 0$ incorporates the additional end-of-sequence token. (Bisani and Ney, 2008)

Besides the graphone length limit $L$, which significantly impacts the size of the resulting graphone inventory, there is another external parameter $M$, which stands for the maximum history length. "Together with L it defines the effective span of the model, i.e. the number of letters or phonemes that affect the estimated probabilities at a given position." (Bisani and Ney, 2008).

To prevent overfitting, evidence trimming is used to trim values below a threshold $\tau$ which causes unlikely graphones to gradually die out during the iteration phase. This is done by replacing the evidence for $q$ in Equation (2.14) ($e(q,h;\boldsymbol{\vartheta})$) with the following case:

$$\hat{e}(q,h;\boldsymbol{\vartheta}) = \begin{cases} 0 & \text{if } e(q,h;\boldsymbol{\vartheta}) < \tau \\ e(q,h;\boldsymbol{\vartheta}) & \text{otherwise} \end{cases} \tag{2.16}$$

The threshold $\tau$ is adjusted on validation data. (Bisani and Ney, 2008)

Similar to typical n-gram language models, the estimation equation (Equation (2.14)) faces a modeling problem: "[A]ny graphone that can be construed from the training examples will receive some probability mass, whereas only a small subset of these is expected to contribute to the 'correct' model." (Bisani and Ney, 2008). Effective smoothing

techniques, especially Kneser-Ney[1] smoothing, have proven to be important for dealing with this issue. First, absolute discounting and interpolation is applied to the evidence for $q$ equation (Equation (2.14) on the previous page): (Bisani and Ney, 2008)

$$p_M(q|h;\vartheta') = \frac{\max\{e(q,h) - d_M, 0\}}{\sum_{q'} e(q',h)} + \lambda(h) p_{M-1}(q|\bar{h}) \tag{2.17}$$

where $M$ is an added subscript to indicate the order of distribution, $d_M \geq 0$ is a discount parameter and $p_{M-1}(q|\bar{h})$ is the generalized, lower order $(M-1)$-gram distribution conditioned on the reduced history $\bar{h}_i = (q_{i-M+2}, ..., q_{i-1})$. $\lambda(h)$ is used for having the overall distribution sum become 1. (Bisani and Ney, 2008)

Because evidence values can become even smaller than the discount, graphones with evidence values below the discount parameter are excluded from the model as an additional form of evidence trimming. In contrast to the evidence trimming in Equation (2.16) on the preceding page, the discounted evidence is distributed over unseen events instead of seen events. (Bisani and Ney, 2008)

As a next step, back-off distribution $p_M - 1$ following the marginal preserving approach by Kneser and Ney (1995) is incorporated in Equation (2.17) to fully apply Kneser-Ney smoothing. A consistency constraint is imposed for all reduced histories $\bar{h}$: (Bisani and Ney, 2008)

$$\sum_{h \in \bar{h}} p_M(q|h) \sum_{q'} e(q',h) = \sum_{h \in \bar{h}} e(q,h) \tag{2.18}$$

Finally, the evidence for $q$ equation with full Kneser-Ney smoothing is obtained by combining Equation (2.17) and Equation (2.18). Solving for $p_{M-1}(q|\bar{h})$ under the constraint that $p_{M-1}$ is smoothed as well, results in the final formula: (Bisani and Ney, 2008)

$$p_{M-1}(q|\bar{h}) = \frac{\hat{e}(q,\bar{h})}{\sum_{q'} \hat{e}(q',\bar{h})} \tag{2.19}$$

with $\hat{e}$ defined as the reduced evidence

$$\hat{e}(q,\bar{h}) := \sum_{h \in \bar{h}} \min\{e(q,h), d_M\}. \tag{2.20}$$

The training iteration is started by initializing the unigram model with flat probability

[1] Kneser-Ney smoothing refers to absolute discounting with interpolation and a marginal preserving back-off distribution; first applied by Kneser and Ney (1995)

distribution (see Equation (2.15) on page 15) for all multigrams to have the same probability. In addition, an alternative initialization counts how often a graphone potentially occurs in each word in the training set regardless of overlap with neighboring graphones but following graphone length constraints: (Bisani and Ney, 2008)

$$c(q) := \sum_{i=1}^{N} \sum_{l_1=1}^{|g_i|} \sum_{l_2=l_1}^{|g_i|} \sum_{r_1=1}^{|\varphi_i|} \sum_{r_2=r_1}^{|\varphi_i|} \times \delta((g_{l_1} \smile \ldots \smile g_{l_2}, \varphi_{r_1} \smile \ldots \smile \varphi_{r_2}) = q) \tag{2.21}$$

With these counts, the initial probability distribution is computed by applying the first part of Kneser-Ney smoothing (see Equation (2.17) on the previous page). Then, higher order $M$-gram models are created using the $(M-1)$-gram model from Equation (2.19) on the preceding page. This only allows histories which correspond to $M$-grams that still exist after discounting in the lower order model. (Bisani and Ney, 2008)

The data used to optimize the discount values has to be separate from the data used to calculate the evidence values because using the same data would result in an underestimation of the discount values. Therefore, the training set $\mathcal{O}$ is separated into a training set $\mathcal{O}_t$ which is used for calculating the evidence values and a smaller held-out set $\mathcal{O}_h$ which is used to adjust the discount parameters. The EM algorithm is used for the training process. Using the normal EM algorithm, however, will lead to overfitting since it will strictly improve the likelihood of the training set $\mathcal{L}(\mathcal{O}_t)$ in each iteration and will cause the likelihood of the held-out set $\mathcal{L}(\mathcal{O}_h)$ to start decreasing. Because of that, a discounted EM algorithm is used where the discount values are updated (see Listing 2.1). (Bisani and Ney, 2008)

Finally, after the model has been estimated, the finished G2P model can be used to transcribe (unseen) grapheme sequences to their respective pronunciations. Usually, the most likely transcription is used as the resulting phoneme sequence by searching for the

**Listing 2.1:** Discounted EM algorithm for the training process of the Sequitur G2P model in pseudocode. (Bisani and Ney, 2008; edited to match equation order)

```
for M = 1 to M_max:
    initialize M-gram model with (M−1)-gram model
        p_M(q|h) = p_{M−1}(q|h̄)
    initialize the additional discount parameter
        d_M = d_{M−1}
    repeat until L(O_h) stops increasing:
        compute evidence according to Equation (2.13)
        if L(O_h) did not increase:
            adjust discount parameters d_1, ..., d_{M−1} by direction set method
                d = argmax_{d'} L(O_h; d')
        update model according to Equation (2.17) and Equation (2.20)
```

| Author | PER(%) | WER (%) |
|---|---|---|
| Galescu and Allen (2002) | 7.0 | 28.5 |
| Chen (2003) | 5.9 | 24.7 |
| Bisani and Ney (2008) | **5.88 ± 0.18** | **24.53 ± 0.65** |

**Table 2.2:** Results of Sequitur G2P (bottom row) in comparison to other G2P models for the CMUdict dataset. (Bisani and Ney, 2008)

most likely graphone sequence that matches the spelling:

$$\varphi(g) = \varphi\left(\underset{q \in Q^* | g(q)=g}{\operatorname{argmax}} p(q)\right) \tag{2.22}$$

Bisani and Ney (2008) compared their model to other G2P systems at that time and showed better or equal results. The results for CMUdict are shown in Table 2.2. Sequitur G2P has been well-established and is still being used for state-of-the-art ASR models (Milde et al., 2017). The code is published on GitHub under the GNU Public License. Sequitur G2P is used at Fraunhofer IAIS and hence will also be used in this work.

### 2.1.2 Sequence-to-Sequence G2P Conversion

Sequences are a challenge for DNNs because the inputs and outputs are of unknown dimensions. To solve this problem for the machine translation field, Sutskever et al. (2014) proposed the LSTM architecture, an extended version of a traditional Recurrent Neural Network (RNN) that was first introduced by Hochreiter and Schmidhuber (1997), to build a Seq2Seq model.

RNNs are able to look at recent information to help performing the current task. In a language model, for example, it is able to predict the last word in the sentence "A hand has five *fingers*" easily because the context is quite clear. In this case, the gap between the relevant information and the time where this information comes to use is small. RNNs are able to handle this small gap and access the information if needed. However, traditional RNNs do not work well with long-term dependencies. In a larger text where relevant information was mentioned at the beginning, it struggles to connect that information to help make a prediction at a later time. For example, in the sentence "I graduated from medical school in 2009. ... I work as a *doctor*.", the last word is probably an occupation, but the past context of "medical school" is needed to narrow down the prediction. The bigger the gap to previous information becomes, the harder it gets to learn to connect previous information, up to a point where it is entirely unable to do so. (Olah, 2015)

In contrast to a traditional RNN, LSTM models are specifically designed to deal with

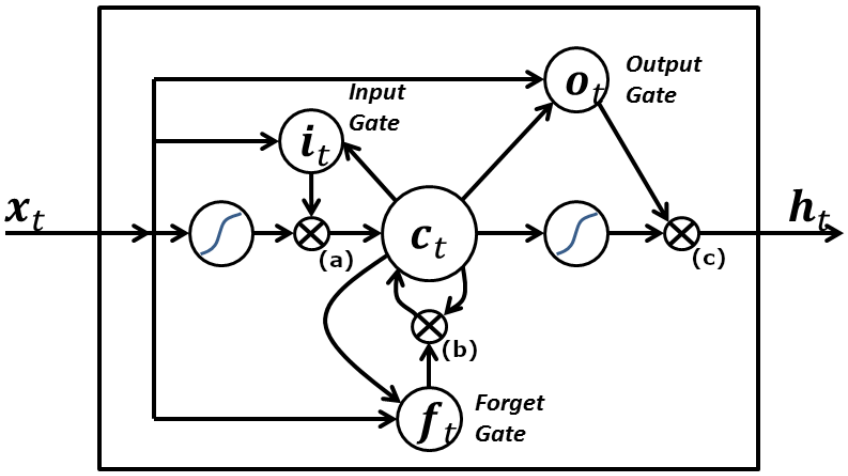


**Figure 2.2:** Architecture of an LSTM cell (adapted from Wikimedia Commons, 2015).

long-term-dependencies, making them able to remember information for long periods of time. They have a chain structure of repeating cells of neural networks, like a traditional RNN, but the cells themselves have a more complex architecture. (Olah, 2015)

Figure 2.2 shows an LSTM cell. The core of an LSTM cell is the cell state $c_t$. It transports information straight through the cell, having a few element-wise multiplications $\otimes$ that can change the information. The first layer of an LSTM cell is a sigmoid layer called *forget gate* $f_t$. Depending on the values of the output of the previous LSTM cell $h_{t-1}$ and the current input $x_t$, it controls for each number in the previous cell state $c_{t-1}$ what information should be kept (1) and what should be forgotten (0). The second layer is another sigmoid layer called *input gate* $i_t$ that decides which values in $c_t$ should be updated. The third layer is an activation layer that decides new candidates that could be added to $c_t$, storing them into a vector $\bar{c}_t$. After passing those three layers, the old cell state $c_{t-1}$ will be updated to $c_t$ in $\otimes$ a by multiplying it by $f_t$ to forget the unimportant information, and adding $i_t * \bar{c}_t$ to add new candidates depending on how much $i_t$ decided to update each value in $\otimes$ b. Finally, $c_t$ goes through the last sigmoid layer called *output gate* $o_t$ which decides what parts of $c_t$ are kept to be transported to the next LSTM cell. To generate the actual output, $c_t$ first goes through tanh to get values between $-1$ and 1, and is then multiplied by $o_t$ in $\otimes$ c to only extract the information the output gate has decided to. (Olah, 2015)

In their implementation, Sutskever et al. use an LSTM network as an encoder to obtain a fixed dimensional vector representation of an input sequence, and another LSTM network that is conditioned on the input sequence as a decoder to extract the output sequence from the vector (Sutskever et al., 2014). This way, a model can be trained that maps source language input sequences to target language output sequences.

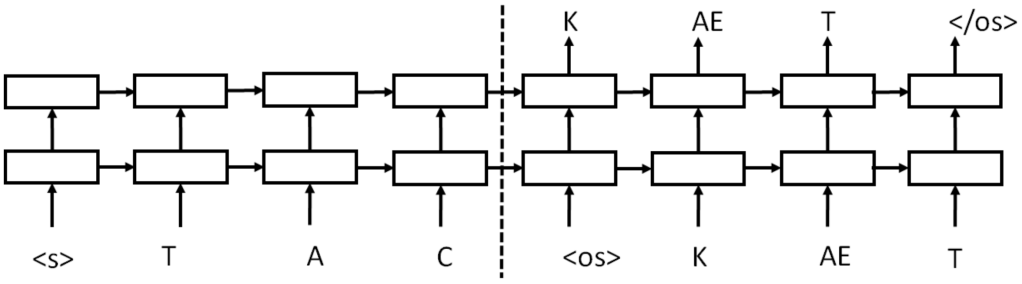


**Figure 2.3:** G2P Encoder-Decoder-LSTM model representation for the grapheme sequence "C A T" with reversed input sequence and phoneme output in ARPABET notation. (Yao and Zweig, 2015)

While experimenting, Sutskever et al. (2014) found out that reversing the order of the input sequence elements positively influences the performance:

> So for example, instead of mapping the sentence $a$, $b$, $c$ to the sentence $\alpha$, $\beta$, $\gamma$, the LSTM is asked to map $c$, $b$, $a$ to $\alpha$, $\beta$, $\gamma$, where $\alpha$, $\beta$, $\gamma$ is the translation of $a$, $b$, $c$. This way, $a$ is in close proximity to $\alpha$, $b$ is fairly close to $\beta$, and so on, a fact that makes it easy for SGD[1] to "establish communication" between the input and the output. We found this simple data transformation to greatly boost the performance of the LSTM.
>
> **Sutskever et al. (2014)**

If the LSTM reads the input sentence in reverse, many short term dependencies in the data are introduced that make the optimization problem much easier (Sutskever et al., 2014).

In 2015, Yao and Zweig successfully applied Sutskever et al.'s method to the G2P task, giving similar results than traditional joint-sequence models. A side-conditioned generation model (encoder-decoder LSTM) as well as two alignment-based models (uni-directional LSTM & bi-directional LSTM) including variations were implemented and compared.

Yao and Zweig's Encoder-Decoder-LSTM directly follows the method proposed in Sutskever et al. (2014), using a depth of two layers for their LSTM model. Figure 2.3 shows the two-layered model. The encoder LSTM (left) reads the reversed input grapheme sequence "⟨s⟩ T A C", where ⟨s⟩ indicates the beginning of the sequence. After the last hidden layer activation, the decoder LSTM (right) is initialized. It produces "⟨os⟩ K AE T" as phoneme prediction of the input sequence and uses "K AE T ⟨/os⟩" as the output sequence. ⟨os⟩

[1] Stochastic gradient descent

| Model | PER(%) | WER (%) |
|---|---|---|
| Sequitur G2P (baseline) | 5.88 | 24.53 |
| Encoder-decoder LSTM | 7.53 | 29.21 |
| Encoder-decoder LSTM (2 layers) | 7.63 | 28.61 |
| Uni-directional LSTM | 8.22 | 32.64 |
| Uni-directional LSTM (window size 6) | 6.58 | 28.56 |
| Bi-directional LSTM | 5.98 | 25.72 |
| Bi-directional LSTM (2 layers) | 5.84 | 25.02 |
| Bi-directional LSTM (3 layers) | 5.45 | 23.55 |

**Table 2.3:** Results on the CMUdict dataset including Sequitur G2P baseline. (Yao and Zweig, 2015)

and ⟨/os⟩ indicate the start and end of the output sequence respectively. (Yao and Zweig, 2015)

The encoder LSTM represents the entire input sequence in the hidden layer activities which are used as the initial activities of the decoder. The decoder LSTM works as a language model. It uses the past phoneme sequence to predict the next phoneme. The decoder stops predicting after outputting ⟨/os⟩.(Yao and Zweig, 2015)

The encoder-decoder LSTM was trained using 500 dimensional projection and hidden layers. Yao and Zweig used "back-propagation through time (BPTT), with the error signal originating in the decoder network" (Yao and Zweig, 2015). Beam search was used to generate the phoneme sequence during decoding, selecting the hypothesis sequence with the highest posterior probability as the decoding result. The batch size per iteration was set to 1 for the CMUdict data as it performed best with this mini-batch on the validation data. The the order of the training sequences was randomly permuted in each epoch. The initial learning rate was set to 0.007 and halved throughout training if the validation loss did not improve. (Yao and Zweig, 2015)

Yao and Zweig (2015) also tested two approaches with relaxed constraints by adding an explicit alignment to the sequence translation. The first approach is an uni-directional LSTM that adds a reference to the past phoneme prediction. This makes the current phoneme prediction dependent on both the letter sequence and the phoneme predictions from the sequence beginning. The second approach is a bi-directional LSTM which uses two RNNs to process the input sequence from left-to-right and right-to-left respectively and then combines their outputs. Here, both RNNs depend on the letter sequence, but only the forward direction LSTM is dependent on the past phoneme predictions. (Yao and Zweig, 2015)

Table 2.3 on the previous page shows the results of their models in comparison to the Sequitur G2P baseline, tested on the CMUdict dataset. The encoder-decoder LSTM models as well as the uni-directional LSTM models perform a bit worse than the baseline. For the bi-directional LSTM, the two-layered model shows improvements in the PER and the three-layered model even significantly exceeds the baseline results in both PER and WER. (Yao and Zweig, 2015)

## 2.2 Wiktionary as Source for Pronunciations

Wiktionary[1] is a collaborative project by Wikimedia Foundation to create a free, multilingual online dictionary. It currently consists of 33,412,117 articles, including 6,335,242 English and 906,444 German ones. An article is almost identically structured across all languages and consists of lexical information, including pronunciations and etymologies. Pronunciations are transcribed in IPA symbols, which is a unified form of representing oral sounds, created by the International Phonetic Association (IPA). Wiktionary is commonly used as a data source in the ASR field, e.g. by Milde et al. (2017) and Sokolov et al. (2019).

Schlippe et al. (2014) analyzed how Wiktionary performs as a source for building pronunciation dictionaries for ASR usage. The pronunciation quality and quantity of the *GlobalPhone* database, a rule-based pronunciation dictionary manually cross-checked by professionals, was compared to the pronunciations found in Wiktionary by building a G2P model on those two data sources. Table 2.4 on the following page shows the word error rates for all created models. While the GlobalPhone models generally performed best, the results for the German Wiktionary models did not differ that much. This implies that Wiktionary seems to be a consistent and reliable data source for the German language.

Because the authors experienced bad phoneme error rates for the resulting English G2P dictionary in an earlier study which were explained by “a difficult g2p correspondance [sic] and corrupted training material from Wiktionary” (Schlippe et al., 2012), several filtering mechanisms were additionally applied to automatically identify and reject flawed or inconsistent pronunciations. As seen in Table 2.5 on the next page, the filters helped improve the WERs, although they perform differently across the tested languages. The authors address the need to further investigate better methods of handling flawed or inconsistent Wiktionary pronunciations.

[1] https://www.wiktionary.org/

| | cs | de | en | es | fr | pl |
|---|---|---|---|---|---|---|
| GlobalPhone baseform | 15.59 | 16.71 | 14.92 | 12.25 | 20.91 | 15.51 |
| GP 1-best | 17.58 | 16.50 | 18.15 | 12.59 | 22.68 | 15.78 |
| wikt 1-best | 18.72 | 16.81 | 28.86 | 12.82 | 25.79 | 17.21 |
| GlobalPhone with variants | 15.62 | 17.11 | 11.52 | 11.97 | 20.41 | 14.98 |
| GP n-best | 18.06 | 17.06 | 18.66 | 12.32 | 22.68 | 15.68 |
| wikt n-best | 19.32 | 17.40 | 37.82 | 12.81 | 25.17 | 17.34 |
| Grapheme-based | 17.56 | 17.83 | 19.15 | 14.06 | 23.36 | 15.38 |

**Table 2.4:** Word error rates (%) of ASR systems with *GlobalPhone* & Wiktionary G2P-based dictionaries. (Schlippe et al., 2014)

| | cs | de | en | es | fr | pl |
|---|---|---|---|---|---|---|
| GP 1-best | 17.58 | 16.50 | 18.15 | 12.59 | 22.68 | 15.78 |
| wikt 1-best | 18.72 | 16.81 | 28.86 | 12.82 | 25.79 | 17.21 |
| wikt $G2P$ | 17.86 | 17.18 | 30.00 | 13.14 | 25.62 | 17.00 |
| wikt $Len$ | 18.24 | 17.13 | 23.68 | 13.50 | 25.48 | 17.38 |
| wikt $G2P_{Len}$ | 17.85 | **16.79** | 24.74 | 13.05 | 25.59 | 17.31 |
| wikt $Eps$ | **17.74** | 17.12 | 22.85 | 12.99 | **23.19** | 16.98 |
| wikt $G2P_{Eps}$ | 18.15 | 17.08 | 22.90 | 12.86 | 25.44 | 16.68 |
| wikt $M2NAlign$ | 18.20 | 17.53 | **20.97** | **12.25** | 25.70 | 16.87 |
| wikt $G2P_{M2NAlign}$ | 17.93 | 17.18 | 23.73 | 13.64 | 25.03 | **16.57** |
| Grapheme-based | 17.56 | 17.83 | 19.15 | 14.06 | 23.36 | 15.38 |

**Table 2.5:** Word error rates (%) using filtered Wiktionary G2P-based dictionaries, highlighting the best result for each language. (Schlippe et al., 2014)

## 2.3 Phoneme Mapping for Non-Native Pronunciation Variants

Foreign accents of non-native speakers can be a challenge in ASR. Depending on the proficiency of the speaker in the target second language (L2), the first language (L1) phoneme set will be (partly) used to pronounce foreign words. An approach to solve this challenge is to map the L2 phoneme set to the L1 phoneme set. From 2000 to 2004, Stefan Schaden conducted a project about lexical modeling of foreign accents at Ruhr-Universität Bochum where he addressed the pronunciation differences of non-native speakers. He developed a rule-based system which maps the phonemes of a canonical pronunciation to those used in the speaker's native language, varying between four Accent Levels (AL) depending on the severity of phoneme changes (Schaden, 2003). While AL 1

only contained some minor allophonic deviations, AL 4 almost fully transferred the L1 phoneme set to the L2 one. In 2006, Schaden extended his research by adding a *Phonetic Distance Measure* which is calculated by the following two components (Schaden, 2006):

- **Edit Distance**: Measure to identify the minimum number of manipulations (substitutions, insertions and deletions) needed to transform sequence A into sequence B using the *Levenshtein distance*.
- **Phonetic Segment Similarity**: Measure that assigns a weight factor to substitution operations in order to account for the similarity of individual phonetic segments.

In a practical example using manual phoneme mapping, Huang et al. (2020) created an Indian English pronunciation dictionary by considering the non-native pronunciation variants. The authors determined an Indian English phoneme set and transformed CMUdict into an Indian English common-word list based on the variation features of Indian English and the phoneme set of American English. Tested on spontaneous Indian English speech clips, the WER was lowered from 22.30 % to 18.82 % by using the Indian English common-word list instead of CMUdict. (Huang et al., 2020)

In contrast to a manual approach that needs professional linguistic and phonetic knowledge, machine learning has been used to build a phoneme mapping system. Goronzy et al. (2004) generated English pronunciations for German words with an English phoneme recognizer to train decision trees. The decision trees were used for predicting the respective English-accented variant from the German canonical transcription. (Goronzy et al., 2004)

In a more recent machine learning approach, Patel et al. (2018) used an *Acoustic Coupling Method* to generate phoneme variants of the source language for a target language. A data set in the source language is used as basis. Additionally, the pronunciations from the source data set are synthesized using a TTS to generate audio clips. The audio is specifically generated based on the noted pronunciations from the data set instead of using standard resources to guarantee that the exact pronunciation is presented in the audio. To generate the target languages pronunciations, a *Pronunciation Learning System* (PLS) based on Bruguier et al. (2017) was built. In contrast to Goronzy et al. (2004), it uses both a pronunciation and an acoustic model for generating the target languages phoneme adaptations as the grapheme sequences also represent valuable cues for learning a realistic pronunciation representation.

The PLS of Bruguier et al. uses a transcript FST with included word FSTs. The transcript FST creates one or more transcripts for each possible verbalized transcript (e.g. “15”, “fifteen”, “one five”). It then creates an FST branch for each verbalized transcript that consists of a sequence of word FSTs and an epsilon-to-word FST (see Figure 2.4 on the

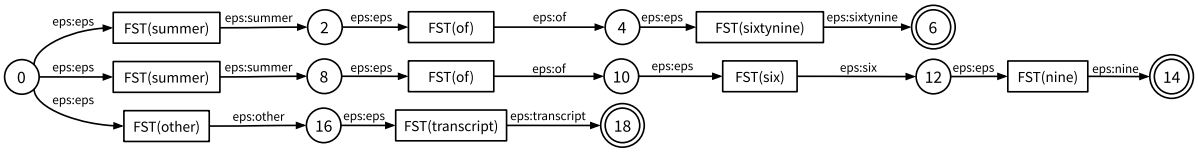


**Figure 2.4:** Transcript FST with included word FSTs (boxes) for the transcript "summer of 69". (adapted from Bruguier et al., 2017)

following page). For a word FST, Bruguier et al. propose six different building methods that were compared against each other, including the *Neural Language Model Expansion FST* which is used in Patel et al.s approach. The *Neural Language Model Expansion FST* is based on the Seq2Seq G2P model by Toshniwal and Livescu (2016), but instead of only finding the best path from the start to end of a sequence, the whole FST is kept as the weights of each arc are further used by combining them with the weight of the acoustic model during the decoding stage. By doing this, Bruguier et al. can control the influence of the acoustic model versus the pronunciation model (G2P) with two hyperparameters:

1. The Boltzmann temperature $\tau$ of the softmax output controls the spikiness and candidate diversity of the output:

$$p_i = \frac{e^{z_j/\tau}}{\sum_j e^{z_j/\tau}} \tag{2.23}$$

   Depending on the value for $\tau$, the following can be achieved:

   - $\tau = 1$: The original model's outputs are preserved.
   - $\tau \to 0$: The softmax output becomes a pure maximum function.
   - $\tau \to \infty$: Each transition becomes equally probable.

2. The beam search algorithm is used to minimize the total cost of a path. The total cost is defined as the sum of the AM and PM costs. By adding a relative weight $\alpha$, the importance of the acoustic versus the pronunciation model can directly be controlled:

$$\text{TotalCost} = \alpha.\text{AMCost} + (1-\alpha).\text{PMCost} \tag{2.24}$$

   Depending on the value for $\alpha$, either the AM or the PM cost is taken more into account, influencing the decision of the beam search.

By using the PLS, Patel et al. generate the pronunciations for the target language based on the source languages pronunciation and audio data. Having source and target language pronunciation pairs, alignments between the different phonemes are found, allowing a

one-to-many source to target phoneme alignment with a range of 0–2 to properly handle diphthongs and consonant pairs:

$$q = (s, \boldsymbol{t}) \in \left( S \times \bigcup_{i=0,1,2} T_i \right) \tag{2.25}$$

where $s$ is the source phoneme, $\boldsymbol{t}$ is the target phoneme sequence, $S$ is the set of source phonemes, $T$ is the set of target phonemes and $T_i$ is the set of elements of T of length $i$. (Patel et al., 2018)

The alignments between $s$ and $\boldsymbol{t}$ are defined as follows:

$$A(s, \boldsymbol{t}) = q_1, ..., q_n \in q * | s_1, ..., s_n = \boldsymbol{s}; \boldsymbol{t_1}, ..., \boldsymbol{t_n} = \boldsymbol{t} \tag{2.26}$$

where an alignment $q_i = (s_i, \boldsymbol{t_i})$. Afterwards, the EM algorithm is used on an observation set to estimate values for $p(q)$ that optimize the likelihood of the training data. (Patel et al., 2018)

The final mapping is defined as follows:

$$\underset{S \to T}{\text{mapping(s)}} = \begin{cases} s, & \text{if } s \in T \\ \text{argmax}_{\boldsymbol{t}}\, p(s, \boldsymbol{t}), & \text{otherwise} \end{cases} \tag{2.27}$$

The source phoneme sequence is not changed if it already exists in the target language's phoneme inventory. If it does not, the target phoneme sequence with the highest probability is used. (Patel et al., 2018)

The results of the automated phoneme mapping were compared to manual mappings with the same restrictions. While the mapping differences from English phonemes to French were 30 %, the differences to German were only 10 %. If no ground truth target language pronunciations or linguistic knowledge is available, Patel et al. conclude that acoustic coupling mapping is comparable to human-generated mapping. (Patel et al., 2018)

## 2.4 Multitask Learning

First introduced by Caruana (1993), multitask learning (MTL) is an approach of modeling the human concept of inductive transfer to a machine learning model. When a human is confronted with a new problem, they use the skills and information they already learned for related problems in the past. As opposed to single task learning where every task is learned separately from each other (see Figure 2.5a on the next page), MTL allows

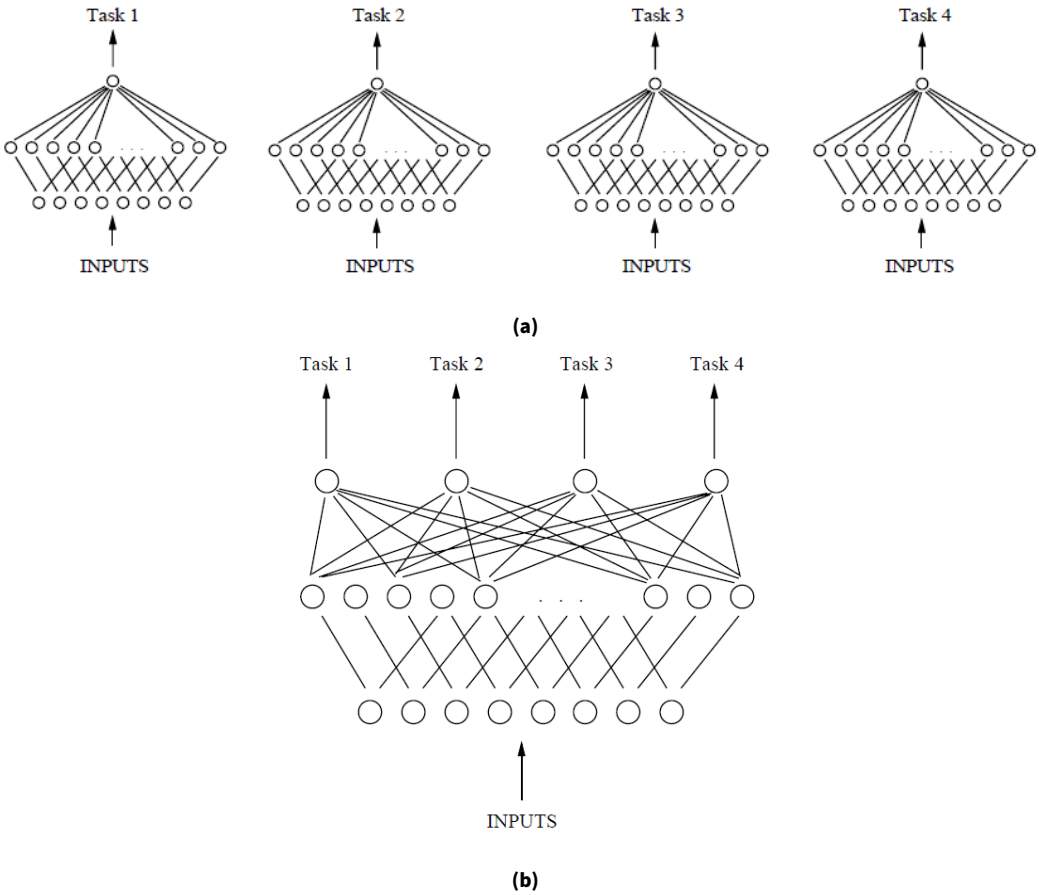


**Figure 2.5:** Single task learning (a) and MTL (b) of four tasks with the same input. (Caruana, 1997)

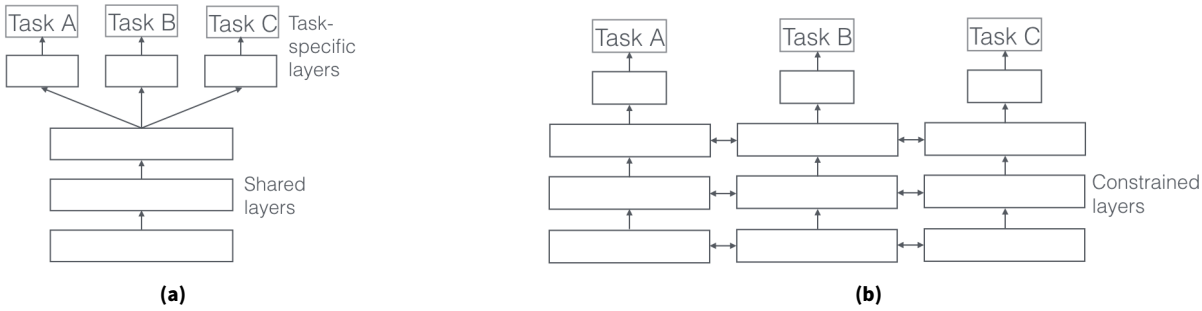


**Figure 2.6:** Hard (a) and soft (b) parameter sharing for multiple tasks. (Ruder, 2017)

learning multiple tasks in parallel by using shared layers and exchanging parameters with each other (see Figure 2.5b).

In practice, MTL can be applied to a DNN by either hard or soft parameter sharing of hidden (input) layers:

- **Hard parameter sharing** (Figure 2.6a)
  Hard parameter sharing is the most common approach for MTL which was first used by Caruana (1993). The input layers are shared between all tasks while the output

layers are kept task-specific. It was found that hard parameter sharing reduces the risk for overfitting because the model is forced to find a more general representation that fits to all tasks instead of only one. (Ruder, 2017)

- **Soft parameter sharing** (Figure 2.6b on the preceding page)
  In soft parameter sharing, the input layers are not generally shared as it is done in hard parameter sharing. Each task has its own input layers, but the distance between their parameters is being regularized so the parameters become more similar to each other. (Ruder, 2017)

The main prerequisite for an MTL model is that the tasks have to be related to each other. "MTL is a method designed for inductive transfer, and inductive transfer between unrelated tasks does not seem sensible" (Caruana, 1997, p.69). An example for a successful use of related tasks is the Multitask Question Answering Network by McCann et al. (2018) that i.a. uses translation, summarization, sentiment analysis and semantic role labeling tasks in a single DNN, all based on the same input layers. Related tasks provide an inductive bias which makes the model learn more general representations. This machine representation of inductive transfer brings several advantages. As already mentioned in the hard parameter sharing description, the risk of overfitting decreases. The data-dependent noise is ignored because a more general representation has to be learned by the model to fit multiple tasks. Also, having multiple tasks helps to differentiate between relevant and irrelevant features since the other tasks provide evidence for the feature's relevance. This shifts the model's focus on those features that are actually important. The resulting biased model prefers representations that other tasks also prefer which makes it possible to introduce new tasks from the same environment in the future while still performing well. (Ruder, 2017)

Overall, MTL seems to fit tasks in the field of natural language processing well since (written and spoken) text contains various cues that can be helpful for multiple tasks simultaneously. MTL has also been applied to the task of G2P conversion: In their approach of multilingual speech recognition, Milde et al. (2017) built a multilingual Seq2Seq G2P model utilizing MTL, also comparing it to monolingual Seq2Seq and Sequitur G2P models. The monolingual models were trained on the German *PHONOLEX* dataset. The multilingual MTL model was simultaneously trained on a German and English G2P task using the *PHONOLEX* and *CMUdict* data sets. They also added a variation of the MTL model with a multi-alphabet approach by adding IPA pronunciations. For this, they generated their own German and English dictionaries by taking the IPA pronunciations from Wiktionary with only American English variants for the English pronunciations to match the CMUdict dataset. (Milde et al., 2017)

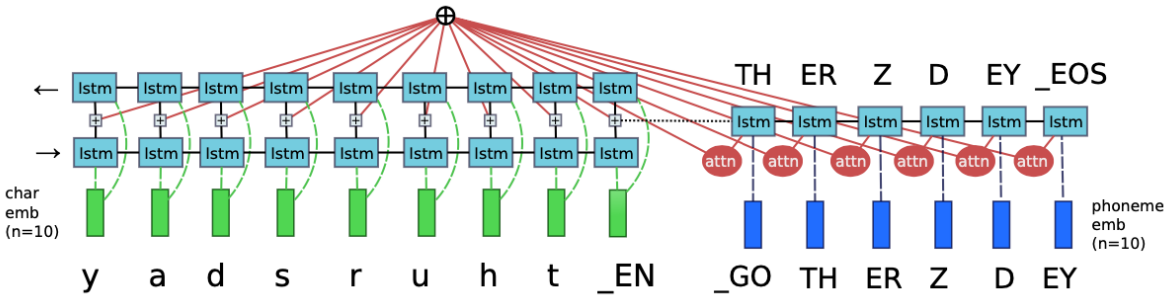


**Figure 2.7:** Seq2Seq G2P model with reversed input and additional language marker for the subtask at the start of each word. (Milde et al., 2017)

Figure 2.7 shows the MTL G2P model by Milde et al. (2017). The model uses character embeddings as encoder inputs and phoneme embeddings as decoder inputs with an embedding size of $n = 10$. A language marker is added at the start of each input sequence for classifying the source language. Also, the input character sequences are reversed. The encoder is built as a stacked bi-directional LSTM to represent past and future dependencies. One LSTMs reads the input sequence forwards while the other one reads it backwards. The output of both LSTMs is combined with a vector sum. Also, a simple residual connection is added between the two LSTM layers to make it easier to learn the identity function by bypassing the layers computation. This is done by additionally passing the output of the first LSTM layer to the second one. To facilitate the information flow from source to target sequence, global attention is added to the decoder. For this, an attention vector is created at each phoneme generation step that holds the weighted sum over all hidden states of the encoder. Beam search is used to generate a softmax distribution over the output vocabulary at each decoding step. (Milde et al., 2017)

Various models were tested and compared against each other. The results are shown in Table 2.6 on the following page. While the Sequitur G2P model (2) performs best among the single models, Milde et al. also tested two system combinations with this particular model combined with two different MTL models. While the combination with the SAMPA/IPA approach (11) performed best, both combinations outperformed model (2). (Milde et al., 2017)

Table 2.7 on the next page shows the results of selected models that were also tested on specific word groups inside the German PHONOLEX test set, including a set of English loan words (anglicisms). Surprisingly, the monolingual Sequitur G2P model performed better than both MTL models, even though they additionally contained an English G2P task. (Milde et al., 2017)

| Model | PHONOLEX set | PER (%) | WER (%) |
|---|---|---|---|
| (1) Sequitur G2P (model order 10) | core | **1.98** | 11.54 |
| (2) Sequitur G2P (model order 6) | core | **1.98** | **11.30** |
| (3) Sequitur G2P (model order 6) | full | 5.41 | 29.86 |
| (4) seq2seq-attn (biLSTM 256x3, d=0.5) | full | 6.09 | 32.69 |
| (5) seq2seq-attn (biLSTM 256x3, d=0.5) | core | 2.49 | 13.64 |
| (6) seq2seq-attn (biLSTM res. 256x3, d=0.5) | core | 2.37 | 12.75 |
| (7) seq2seq-attn (biLSTM res. 256x3, d=0.5) + MTL (de/en) | core | 2.57 | 14.12 |
| (8) seq2seq-attn (biLSTM 512x3, d=0.5) + MTL (de/en) | core | 2.41 | 13.32 |
| (9) seq2seq-attn (biLSTM res. 512x3, d=0.5) + MTL (SAMPA/IPA) | core | 2.06 | **11.30** |
| (10) System combination (2)+(6) | core | 1.88 | 10.33 |
| (11) System combination (2)+(9) | core | **1.70** | **9.52** |

**Table 2.6:** PER and WER performance of all tested models for the German PHONOLEX test set. Column "PHONOLEX set" implies which data was used to train the respective model. (Milde et al., 2017)

| Model | PER (%) | WER (%) |
|---|---|---|
| Sequitur G2P (model order 6) | **10.20** | **38.24** |
| seq2seq-attn (biLSTM 256x3) | 16.93 | 52.94 |
| + MTL (de/en) | 17.20 | 61.76 |
| + MTL (SAMPA/IPA) | 12.60 | 47.06 |

**Table 2.7:** PER and WER performance of selected models specifically for English loanwords inside the German PHONOLEX test set (34 words, 2.59 %). (Milde et al., 2017)

# Chapter 3

# Methodology

**Contents**


3.1 Research Questions . . . . . 31
3.2 Data Collection . . . . . 31
3.3 General Evaluation . . . . . 36


This chapter describes the methodology used for this work. After defining the research questions, the sourcing and preparation of the required data are described. Also, the evaluation methods are explained to get an understanding of the metrics needed for interpreting the anglicism recognition results.

## 3.1 Research Questions

The following research questions have been defined for this work:

#### Q1: How can anglicism recognition be improved in German ASR?

Because a German ASR model is trained on the German phoneme set, generating the correct pronunciations for anglicisms will often result in wrong pronunciations since anglicisms mostly use the English phoneme set.

#### Q2: Can pronunciation generation of anglicisms be improved considering their English etymology?

Since anglicisms are of English heritage, utilizing the English language might be a solution for getting the correct phoneme conversions. It has to be investigated if and how the English etymology can help improve the generation of correct anglicism pronunciations.

#### Q3: How can anglicisms be distinguished in the German language?

To potentially treat anglicisms differently when generating pronunciations, they have to be correctly detected. It has to be examined how anglicisms can be distinguished from native German words.

## 3.2 Data Collection

Several resources were needed to conduct experiments for improving anglicism recognition. The following sections describe what data sources have been used and how the data has been prepared for being used in this work.

### 3.2.1 Anglicism Testset

Media data containing anglicisms was collected to create a dedicated anglicism test set. Recordings of the following topics were chosen:

- Releases of the German television news service “Tagesschau” that contained anglicisms
- YouTube videos containing business terms of English heritage
- YouTube videos containing technology related terms of English heritage
- YouTube videos containing anglicisms in colloquial speech

Table A1 on page 113 shows all selected videos including the source links, durations and topic labels. The 15 videos marked with * have already been collected by the team lead Dr. Christoph Andreas Schmidt prior to the start of this work. The other 7 videos have been chosen independently to complement the existing files with further technical terms (e.g. “Venix Tech News” episodes) and colloquial speech containing (inflected) anglicisms (e.g. videos by the YouTuber “Rezo”).

All sections that contained anglicisms were manually detected and transcribed by using the tools *ELAN* and *Simple-ELAN* (Max Planck Institute for Psycholinguistics, The Language Archive, 2020). With the ELAN tools, audio segments can be selected in the oscillogram to annotate them inside a dedicated text field. Speaker names can be specified by using an `@`-notation. Although it was not necessary for this work, the speakers have been declared for the test set since it might be helpful in future projects. Figure 3.1 on the next page shows an example of an annotation in *Simple-ELAN*.

Additionally, the anglicisms were marked as entities in the annotations to enable the possibility for evaluating a specific *Entity Error Rate*. For this, an additional feature has been implemented in the metadata conversion step of the benchmarking process that enables the parsing of entities from the annotation text. The anglicisms could then be marked in the annotation using a specific notation that states the marked word and entity type. An entity notation uses ## as start and end markers. Entities can be declared inside those markers, separating multiple terms with #. For each entity, its type is specified after a semicolon. Listing 3.1 on the following page shows an example containing entity annotations. The anglicism entities for the test set “Anglicisms 2020” have been marked automatically by using a python script that modified the annotations based on an anglicism list that is described in Section 3.2.2 on the next page. A total of 1,362 anglicisms were marked in the test set.

Table 3.1 on page 34 shows the statistics of the resulting test set “Anglicisms 2020”. Overall, 1.31 hours of audio have been used with an average length of 4.39 seconds per

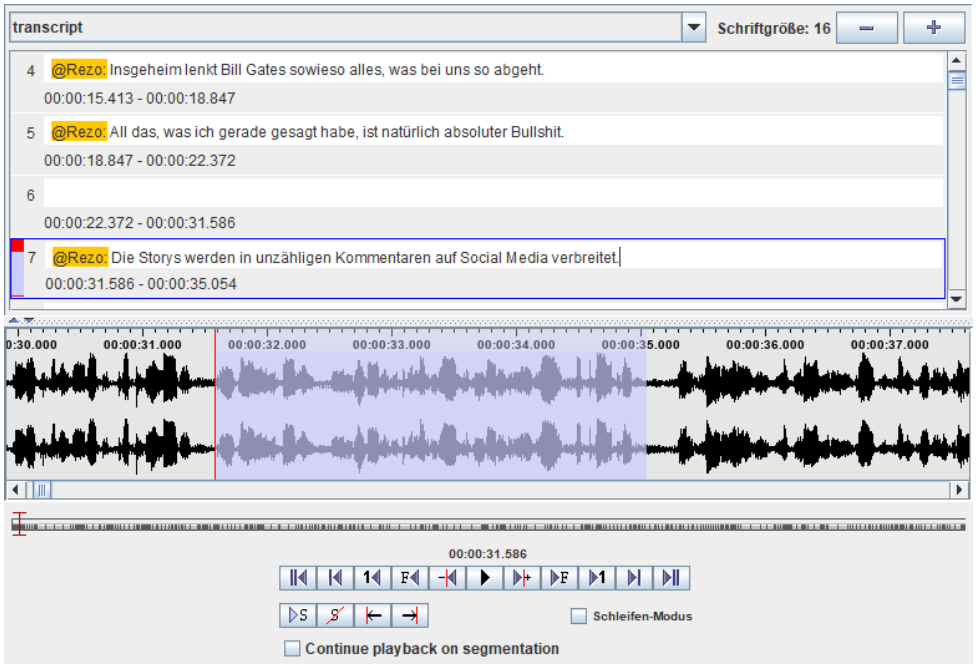


**Figure 3.1:** Manual annotation for the file “Rezo - Die Zerstörung der Presse” in Simple-ELAN.

segment. The test set annotations contain a total of 14,028 words, consisting of a set of 4,660 unique words. In average, a segment contained 13.10 words.

### 3.2.2 Anglicism List

To have a basic set of anglicisms, predefined anglicism lists were sourced from publicly available resources. The first resources were the German *Wiktionary*’s anglicism[1] and pseudo-anglicism[2]indices, containing 4,804 words in total. As a second resource, the anglicism index by *Verein der Deutschen Sprache*[3] (VDS) that contained 7,418 words was used. The VDS anglicism index is a large and well-maintained collection of Anglicisms in German speech that offer descriptions, German equivalents and a specific state that indicates how well-established the respective anglicism is in the German language (Elfers, 2020). Combined, the crawled anglicisms contained 11,839 words.

Based on the combined anglicism list, the German Wiktionary was crawled to obtain

**Listing 3.1:** Entity notation in the annotation of “Rezo - Die Zerstörung der Presse” that marks the anglicisms “Social Media” and “Storys”.

```
@Rezo: Die Storys werden in unzähligen Kommentaren auf Social Media verbreitet. ##Social
    ↪ Media;Anglicism#Storys;Anglicism##
```

[1] https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Anglizismen

[2] https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Anglizismen/Scheinanglizismen

[3] https://vds-ev.de/denglisch-und-anglizismen/anglizismenindex/ag-anglizismenindex

| | |
|---|---|
| **Number of Segments** | 1,071 |
| **Number of Words** | 14,028 |
| **Avg. Words per Segment** | 13.10 |
| **Unique Words** | 4,660 |
| **Overall Length (h)** | 1.31 |
| **Avg. Segment Length (s)** | 4.39 |

**Table 3.1:** Statistics of the test set “Anglicisms 2020”.

| | Person | Wortform | |
|---|---|---|---|
| Präsens | ich | loade down<br>downloade | |
| | du | loadest down<br>downloadest | |
| | er, sie, es | loadet down<br>downloadet | |
| Präteritum | ich | loadete down<br>downloadete | |
| Konjunktiv II | ich | loadete down<br>downloadete | |
| Imperativ | Singular | loade down!<br>downloade! | |
| | Plural | loadet down!<br>downloadet! | |
| Perfekt | Partizip II | | Hilfsverb |
| | downgeloadet<br>gedownloadet | | haben |
| *Alle weiteren Formen:* Flexion:downloaden | | | |

**Figure 3.2:** German inflection box on Wiktionary for the word “downloaden”

additional word inflections. A small number of inflections containing forms of the respective word for a few selected grammatical forms, tenses and persons can be found in a box on the right side of it’s page (see Figure 3.2). Some entries additionally have a dedicated inflection page that contain all available inflections of a word. An inflection page can can be accessed by adding the prefix “Flexion:” to the URL, e.g. `https://de.wiktionary.org/wiki/Flexion:downloaden`. Only the German sections of the Wiktionary pages were used since the inflections should reflect the use of anglicisms in the German language only. Crawling the German word inflections resulted in 5,944 new

anglicisms where 2,821 were taken uniquely from the inflection boxes and 3,123 were taken uniquely from the inflection pages. Combined with the existing anglicism list, it resulted in a total of 17,783 entries. Additionally, a split version of the anglicism list was created that splits words containing whitespaces into their respective parts. The split anglicism list contained 18,967 words. The split list contained 6,312 words that were not already contained in the baseline dictionary.

### 3.2.3 Pronunciation Dictionaries

A pronunciation dictionary is needed to train a G2P model. For the models that were built for this work, the *CMU Pronouncing Dictionary* (CMUdict) in version 0.7b and the *Pronunciation Lexicon PHONOLEX* in its core version were used.

CMUdict is a publicly accessible pronunciation dictionary created by Carneige Mellon University. In version 0.7b, it contains 133,779 English words and their respective pronunciations in American English. The pronunciations are written in ARPABET notation which is considered standard for English pronunciations. The phoneme set consists of 39 different phonemes. This dictionary was chosen because it is freely accessible and widely used in scientific research, e.g. Bisani and Ney (2008), Yao and Zweig (2015), Milde et al. (2017). (Carnegie Mellon University, 2014)

PHONOLEX is a dictionary created by the Bavarian Archive for Speech Signals (BAS). It is a commercial data source and can be used for this work under the scientific license of Fraunhofer IAIS. Its core version only consists of entries that were manually checked by professionals. PHONOLEX core contains 65,427 German words and their respective pronunciations. For the phonemes, an extended SAMPA format (BAS-SAMPA[1]) is used that consists of 52 different phonemes. PHONOLEX core was chosen because it is used for the G2P model that is currently used at Fraunhofer IAIS. (Bavarian Archive for Speech Signals, 2013)

### 3.2.4 Text-to-Speech-Generated Audio Files

For one of the components that were developed for this work, audio files for English pronunciations were needed. To generate the audio files, the text-to-speech services AWS Polly[2] by Amazon and Azure Text-to-Speech[3] by Microsoft were used. Since the needed audio results could be synthesized within the free quota, no fees had to be paid. More information about this data is provided in Section 5.1 on page 52.

[1] https://www.phonetik.uni-muenchen.de/forschung/Bas/BasSAMPA

[2] https://aws.amazon.com/de/polly/

[3] https://azure.microsoft.com/de-de/services/cognitive-services/text-to-speech/

## 3.3 General Evaluation

Three approaches were defined and chosen to be compared against each other according to how well anglicisms were recognized. Each experiment conducted within an approach will result in an anglicism pronunciation dictionary. A monolingual German ASR model will be used as baseline. For the comparison, each anglicism dictionary will be added as a supplementary dictionary in the baseline ASR model. The respective ASR models will be used on three different in-house test sets to generate speech recognition hypotheses. Besides the aforementioned test set "Anglicisms 2020", two German test sets "German Broadcast 2020" and "Challenging Broadcast 2018" have been chosen which represent typical audio data used by clients of the baseline ASR system. Those two test sets are used to ensure that the addition of anglicism pronunciation dictionaries does not compromise the recognition performance of other use cases. The benchmark results will contain the word error rate for all words in general and the entity error rate[1] for anglicisms specifically that were marked in the "Anglicisms 2020" test set. Both error rates can be quantitatively evaluated. Additionally, some samples will be selected to qualitatively show the core differences between the models.

### 3.3.1 Baseline Model

At Fraunhofer IAIS, a new monolingual German ASR model is trained every day based on current data crawls and updates in the contained components to provide up-to-date recognition results for their clients. To ensure that the ASR model used in the experiments is only influenced by the added anglicism pronunciation dictionary, a baseline model was chosen. Since the ASR model "german-default-1.0.124.20201101" (created on November 1st 2020) was used for the first experiment, it was chosen as the baseline ASR model for this work.

### 3.3.2 Anglicism Pronunciation Dictionary

The resulting anglicism pronunciation dictionaries of the different experiments have to be formatted like the pronunciation dictionary in the baseline model. BAS-SAMPA format is used for the phoneme notation. Also, the phoneme set of the baseline model dictionary must be used because adding additional phonemes that were not present in the training stage of the ASR model cannot be processed.

The lines in the baseline pronunciation dictionary are formatted as follows:

**Listing 3.2:** Formatting example of the baseline pronunciation dictionary.

```
Maus    m aU s
```

[1] The evaluation metrics will be explained in Section 3.3.4 on page 39.

The grapheme sequence (⟨Maus⟩) and the phoneme sequence (/m aU s/) are separated by a tab stop or whitespace. Also, in the phoneme sequence, every phoneme is separated by a whitespace. This enables the distinction of diphthongs (e.g. /aI/) and phonemic consonants (e.g. /ts/). The grapheme sequence may not contain whitespaces as it would result in a parsing error.

### 3.3.3 Benchmarking

The benchmarking webservice at Fraunhofer IAIS is a part of the iFinder process that creates ASR results for a predefined test set with a chosen ASR model. The results are saved in a JSON file that shows general statistics (e.g. word error rates) as well as detailed outcomes of the recognition results for each segment in a file. Listing 3.3 on the following page shows an example of a segment result in the JSON file. The result contains i.a. the name of the model (`model`), test set (`testset`) and audio file (`filename`). The main information are the annotated reference (`reference`) and the hypothesis (`hypothesis`) that was recognized by the model. Also, the beginning and end times (in seconds) of the reference (`refStartTimes`, `refEndTimes`) and hypothesis (`hypStartTimes`, `hypEndTimes`) are specified. The hypothesis times are stated in more detail as they reflect the exact timings for all words that have been recognized. Based on the differences in reference and hypothesis, a `path` is generated that describes the differences between reference and hypothesis for each word with the following symbols:

- **M (Match)**: The hypothesis word matches the reference word.
- **S (Substitution)**: The reference word is substituted by a different word in the hypothesis.
- **I (Insertion)**: The hypothesis word is not contained in the reference.
- **D (Deletion)**: The reference word is not contained in the hypothesis.

The JSON benchmark result also contains an entity evaluation for each file that states the total number of entities and the number of correctly recognized entities. With this information, an entity error rate can be calculated. The usage of this metric will be explained in Section 3.3.4 on page 40.

**Listing 3.3:** JSON format: Baseline model benchmark result of segment 3 in “Was ist Machine Learning, eine Einführung - codecentric.AI Bootcamp” from test set “Anglicisms 2020”.


```
{
  "timestamp": "2020-Nov-05 12:04:01",
  "tool": "iFinder",
  "toolVersion": "4.0.3",
  "testset": "Anglicisms 2020",
  "testsetVersion": "1.0.15",
  "experimentName": "",
  "categories": "",
  "parameters": "",
  "accumulationChainName": "",
  "model": "german-default-1.0.124.20201101",
  "modelVersion": "",
  "filename": "\/data\/automatic_deployment\/benchmarking\/anglicisms2020\/media\/
  ↪ y2mate_com_-_Was_ist_Machine_Learning,_eine_Einführung_-
  ↪ _codecentric_AI_Bootcamp_iX9r8wvjKdo_1080p_ganzeSendung.wav",
  "level": "segment",
  "topic": "Speech recognition",
  "segIdx": "3",
  "reference": "schauen wir uns zunächst einmal das Big Picture zum Thema Machine
  ↪ Learning an .",
  "hypothesis": "schauen wir uns zunächst einmal das Big-Picture zum Thema
  ↪ Maschinenreiniger an",
  "path": "MMMMMMDSMMDSMd",
  "refStartTimes": "8590",
  "refEndTimes": "12250",
  "hypStartTimes": "8590 8830 8890 9040 9340 9730 10060 10810 10990 11230 11923",
  "hypEndTimes": "8830 8890 9040 9340 9700 10000 10810 10990 11230 11923 12160"
}
```


In addition to the pure JSON result files, there is a *Benchmark Viewer* that prepares the results from the JSON files in a web interface. This allows for an easier and quicker evaluation of the recognition results. Based on the information in `path`, the Benchmark Viewer marks the differences between reference and hypothesis by coloring them red. Also, the audio can be played to better compare the different results. Based on the start and end times of reference and hypothesis, the exact audio segments are played by clicking a word in the reference or hypothesis text. Figure 3.3 on the following page shows the same segment as shown in Listing 3.3 inside the Benchmark Viewer.

In the hypothesis shown in Listing 3.3 and Figure 3.3 on the next page, a period is missing at the end of the sentence when comparing it to the reference. Even though punctuation differences are marked red in the benchmark viewer, they are not considered as a word

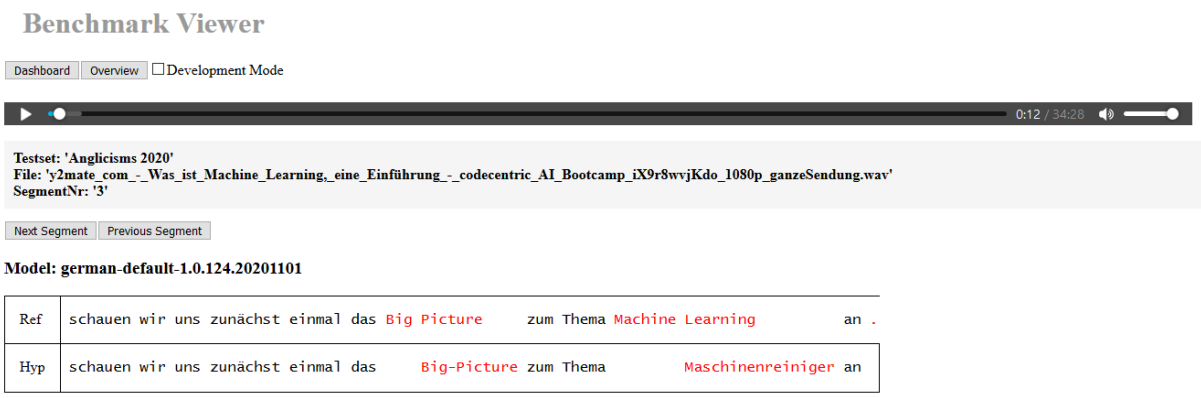


**Figure 3.3:** Benchmark viewer: Baseline model benchmark result of segment 3 in “codecentric.AI Bootcamp - Was ist Machine Learning” from test set “Anglicisms 2020”.

error. Punctuation errors are captured with a separate punctuation error rate. Also, there is a dedicated module in post-processing that handles punctuation in the ASR results. To solely concentrate on the actual word errors that determine the ASR models performance, the punctuation in annotation references was left out in the benchmark examples shown in this work.

### 3.3.4 Evaluation Metrics

To quantitatively evaluate the experiments, three different metrics have been chosen:

#### Word Error Rate (WER)

The WER is used to evaluate the general performance of the ASR and G2P models. It is a metric that quantifies the word errors in a speech recognition hypothesis compared to it’s corresponding reference. Figure 3.4 on the following page shows the resulting WERs of the baseline ASR model for files inside the test set “Anglicisms 2020” inside the benchmark viewer. The lower the WER, the more accurate are the recognition results. For measuring the WER of an ASR model,the Levenshtein distance is used to calculate “the smallest number of insertions, deletions, and substitutions of words required to change the hypothesis sentence into the reference sentence” (Stadtschnitzer, 2018, p.31). Expressed with an equation, the WER is defined as

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + M} \tag{3.1}$$

where $S$ is the number of substitutions, $D$ is the number of deletions, $I$ is the number of insertions, $M$ is the number of matches and $N$ is the number of all words in the reference sequence (Stadtschnitzer, 2018). The paths in the JSON benchmark results (see Listing 3.3 on the previous page) are used to calculate the WER of an ASR model.

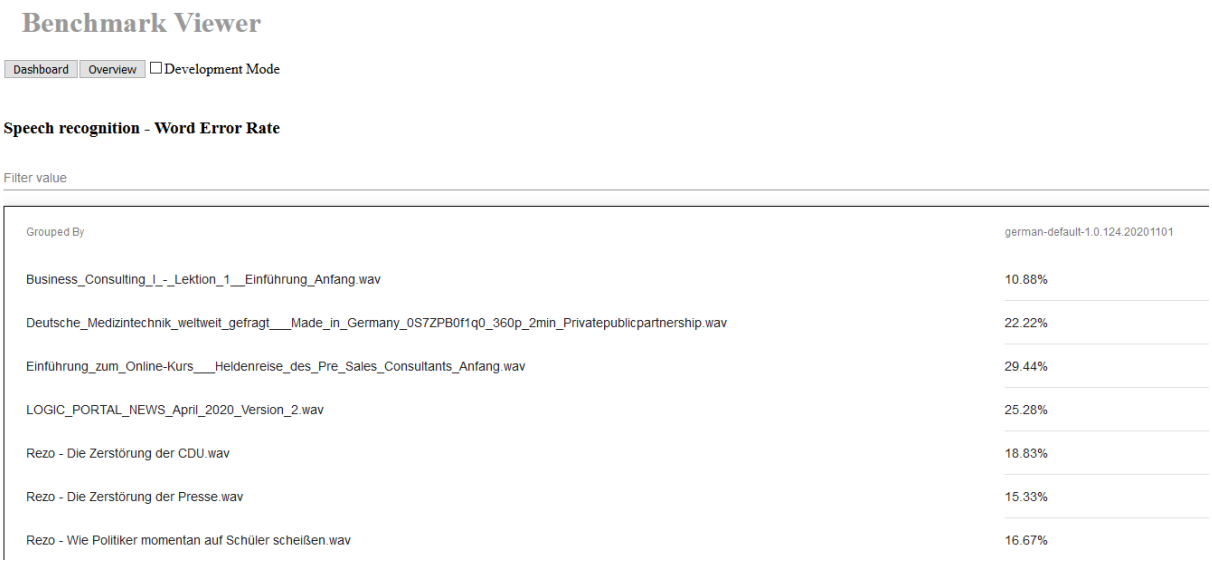


**Figure 3.4:** The Benchmark Viewer at Fraunhofer IAIS showing the WERs of the baseline model for different files in test set “Anglicisms 2020”.

**Sample 3.1:** Segment 3 in “codecentric.AI Bootcamp - Was ist Machine Learning” from test set “Anglicisms 2020”.

| | |
|---|---|
| **reference** | `[...] das Big Picture zum Thema Machine Learning an` |
| baseline | `[...] das Big-Picture zum Thema Maschinenreiniger an` |

For a G2P model, the WER is calculated differently since it transforms grapheme sequences to phoneme sequences instead of recognizing word sequences. Here, the WER is defined as the ratio of pronunciations that contained one or more wrong phoneme in their generated hypotheses.

### Entity Error Rate (EER)

The EER is used to evaluate the performance of anglicism recognition by the ASR model. In the test set “Anglicisms 2020”, 1,362 anglicisms have been marked accordingly. The EER is calculated based on the total number of entities and the number of correctly recognized entities which is provided in the benchmark result. Using this metric, the ASR systems performance on anglicisms can be assessed specifically.

Unfortunately, alternative spellings can cause small irregularities when calculating the WER and EER, especially in the case of compounds. Sample 3.1 shows an example of a recognition result by the baseline model that recognized “Big-Picture” instead of the referenced spelling “Big Picture”. The whitespace has been replaced by a hyphen, but the actual letters and hence the word itself has been correctly recognized. Since cases like this are treated like normal spelling errors by the benchmark process, alternative spellings can result in an increase of both WER and EER. However, all benchmark results are affected

by this phenomenon, so it would only lead to a correct recognition if the pronunciation dictionary contained both “Big” and “Picture” with the matching pronunciations. This would in fact lead to an improvement over the result that recognized the hyphenated spelling because additional required entries would exist in the pronunciation dictionary. Hence, the metrics still reflect an improvement in both word and entity recognition performance and will therefore be used as valid performance measures in this work.

#### Phoneme Error Rate (PER)

The PER is used to evaluate the performance of the G2P model. It is calculated similar to the WER, but using phonemes instead of words. Evaluating the output of a G2P with a predefined test set that contains canonical pronunciations, a lower PER corresponds to a higher accuracy of the model. However, this metric does not distinguish between the similarity of unmatched phonemes. Given the word “Bengel” with its canonical pronunciation `/b E N @ l/`, the pronunciation hypothesis `/b E N E l/` would result in the same PER as `/b E S @ l/`, even though the former sounds much more similar to the canonical pronunciation than the latter does. Therefore, when having to decide for a model based on it’s PER, samples will be looked at in case similar PERs are observed.

# Chapter 4

# Approach 1: Using Pronunciations from Wiktionary

**Contents**


4.1 Data Crawling . . . . . 42
4.2 Dictionary Creation . . . . . 43
4.3 Evaluation . . . . . 43


The German Wiktionary offers pronunciations for most of its entries. Since it is open-source and well-maintained, it has been used as a data resource for various ASR studies, e.g. Milde et al. (2017) and Sokolov et al. (2019). Because everyone is able to add, edit and correct its content quickly and easily, Wiktionary offers a contact point for newly established and colloquial words in the German language that may take some time to appear in a traditional dictionary like Duden due to publication requirements. Wiktionary also provides pre-formatted inflection tables[1] that can be applied to a word by defining its grammatical form which makes it easy to define and maintain newly-added words to the dictionary, resulting in many inflected anglicisms with existing pronunciations.

This approach uses Wiktionary's anglicism pronunciations to create a supplementary anglicism dictionary for the ASR system. Based on the anglicism list (see Section 3.2.2 on page 33), all available pronunciations will be taken from *Wiktionary*.

## 4.1 Data Crawling

As described in Section 3.2.2 on page 33, an anglicism list has been compiled based on two Wiktionary indices that list anglicisms and pseudo anglicisms and the German Anglicism Index by *Verein der Deutschen Sprache* (VDS). Also, inflections of the collected anglicisms have been crawled from the German Wiktionary by looking at the inflection boxes and dedicated inflection pages. The finished anglicism list contained a total of 17,783 words (4,804 from Wiktionary indices, 7,418 from VDS, 5,944 from Wiktionary inflections).

As a last step of data parsing, the German pronunciations of the collected anglicisms were crawled from the German Wiktionary website. 10,918 Wiktionary pages were

[1] https://de.wiktionary.org/wiki/Hilfe:Flexionstabellen

found based on the words in the anglicism list. The crawling of those pages resulted in 9,626 pronunciations for 7,958 words in the anglicism list. For 10,540 anglicisms, no pronunciation could be found.

## 4.2 Dictionary Creation

Some terms from the anglicism list are based on English compounds that contain space characters, e.g. “American Football”. Since the the grapheme sequences should not contain any space characters as it would lead to parsing errors, those terms have to be separated. Luckily, almost all IPA pronunciations for the space-character-containing compounds in the anglicism list were also separated by a space characters and hence matched the corresponding compound element. For 20 terms, the space character had to be added manually to the pronunciation. The terms could then be split automatically by removing the compound entry and creating a new entry for each compound element with its respective pronunciation part.

The retrieved IPA pronunciations for the anglicism list had to be converted to BAS-SAMPA to be compatible with the ASR model. A modified version of the script “phonetics.py”[1] created by Benjamin Milde which is part of his “Speech lex edit” tool (Milde, 2019) was used for the pronunciation conversion. Afterwards, a German dictionary file was created in a format compatible with the pronunciation model used in the ASR system. Listing 4.1 shows the general structure of the resulting pronunciation dictionary. The dictionary contained a total of 9,748 entries. Table A2 on page 114 shows 20 example entries from the anglicism dictionary with their respective pronunciations.

**Listing 4.1:** Extract from the Wiktionary anglicism dictionary

```
downgeloadet    d aU n g @ l O U d @ t
downgeloadet    d aU n g @ l o: d @ t
Downhill        d aU n h I l
Downhills       d aU n h I l s
Download        d aU n l o U d
download        d aU n l O U t
Download        d aU n l o: t
download        d aU n l o: t
```

## 4.3 Evaluation

Since this was the first experiment, it was unknown if the modified pronunciation dictionary would lead to a better or worse performance of the ASR system. Adding pro-

[1] https://github.com/uhh-lt/speech-lex-edit/blob/0abad026eb9afb9fb80f10c48215935601610266/phonetics.py

| Model | Anglicisms 2020 (%) | German Broadcast 2020 (%) | Challenging Broadcast 2018 (%) |
|---|---|---|---|
| baseline | 15.80 | **6.56** | 10.84 |
| wiki-base | **15.72** | 6.57 | **10.83** |

**Table 4.1:** WERs for the baseline and wiki-base ASR models.

**Sample 4.1:** Segment 15 in "Rezo - Die Zerstörung der CDU" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | `[...] und die Armen halt ziemlich abgelost` |
| baseline | `[...] und die Armen halt ziemlich abgeluchst` |
| wiki-base | `[...] und die Armen halt ziemlich abgelost` |

nunciations bears the risk of adding homophones that might cause wrong recognition result depending on the WFST weights. In the case of added anglicisms pronunciations, however, it did not lead to a declined recognition performance as the following results show.

Table 4.1 shows the WERs of the first experiment in the Wiktionary approach (wiki-base) compared to the baseline model. Even though the results only differ slightly, wiki-base shows slightly better results for the test sets "Anglicisms 2020" (Δ 0.08 %) and "Challenging Broadcast 2018" (Δ 0.01 %). For "German Broadcast 2020", the performance dropped a little (Δ 0.01 %). Table A3 on page 115, Table A4 on page 116 and Table A5 on page 117 show the detailed comparison of the results for the specific audio files.

Sample 4.1 shows the benchmark results of segment 15 in "Rezo - Die Zerstörung der CDU" from test set "Anglicisms 2020". It describes a typical example for improvements made by the added anglicism dictionary in wiki-base. While the baseline model recognized the word "abgeluchst" instead of the anglicism "abgelost", wiki-base was able to recognize the correct word. This is caused by the respective pronunciation `/Q a p g @ l u: s t/` for "abgelost " in the pronunciation dictionary of wiki-base which is missing in the baseline dictionary. If a models dictionary does not contain a word needed for the correct recognition, it hence can never be part of any hypothesis the model makes.

Sample 4.2 on the following page shows the benchmark results of segment 0 in "Reportage vom Inneren der Bgm.-Smidt-Brücke (Albrecht, Maike)" from test set "Challenging Broadcast 2018" which also lead to different performances caused by a unique entry in the pronunciation dictionary. The word "Eingangsflurs" was recognized by neither of the

**Sample 4.2:** Segment 0 in "Reportage vom Inneren der Bgm.-Smidt-Brücke" from test set "Challenging Broadcast 2018".

| | |
|---|---|
| **reference** | [...] am Ende des Eingangsflurs eine schwere Eisentür [...] |
| baseline | [...] am Ende des Eingangslagers eine schwere Eisentür [...] |
| wiki-base | [...] am Ende des Eingangs los eine schwere Eisentür [...] |

**Sample 4.3:** Segment 7 in "Rezo - Wie Politiker momentan auf Schüler scheißen" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | Newsflash das ist die Risikogruppe |
| baseline | Newsflashs das ist die Risikogruppe |
| wiki-base | Newsflash das ist die Risikogruppe |

two models as it is not contained in the pronunciation dictionaries. While the baseline model falsely recognized "Eingangslagers", wiki-base recognized the words "Eingangs los". This was most likely caused by an additional entry for the word "los": The pronunciation `/l o: s/` is contained in the baseline dictionary, but only wiki-base has the additional pronunciation `/l u: s/` in its supplementary anglicism dictionary which caused the hypothesis to be different. In this case, both models made a wrong guess for the word "Eingangsflurs". However, if the baseline model had correctly regognized the word, it is possible that the additional entry for "los" would have had a negative impact on the recognition result.

Sample 4.3 shows the benchmark results of segment 7 in "Rezo - Wie Politiker momentan auf Schüler scheißen". In this example, the word "Newsflash" was falsely recognized as "Newsflashs" in the baseline result while it correctly recognized "Newsflash" in the result of wiki-base. Surprisingly, there is no entry for the word "Newsflash" in the supplementary anglicism dictionary of wiki-base. The pronunciations for both words are contained in the baseline dictionary. Since no pronunciations have been added that influenced the result, this outcome can only be caused by differences in the WFSTs of both ASR models.

When building a new ASR model, the WFST paths slightly change if the components are updated. In case of the experiments within this work, the pronunciation dictionary is updated. This can lead to a slightly different pronunciation model WFST since new entries are present. Therefore, changes in the pronunciation dictionary can cause small differences in the WFST of the ASR system. This phenomenon can unfortunately not be prevented.

| Model | Entries | Exclusive Words | Exclusive Pronunciations |
|---|---|---|---|
| wiki-base | 9,748 | 3,122 | 3,651 |
| wiki-v1 | 6,292 | 153 | 195 |

**Table 4.2:** Differences in words and pronunciations in the resulting dictionaries. The exclusive words refer to the words that were contained in the respective model's dictionary, but not in the other models dictionary. Similarly, the exclusive pronunciations refer to the pronunciation variations that were contained in the respective model's dictionary, but not in the other models dictionary.

### 4.3.1 Variation 1: Early Version With Manual Modifications

Since the Wiktionary approach was the first approach that was worked on, an early version of a resulting anglicism dictionary exists that was used in a first test scenario for verifying the general evaluation procedure. In the creation process of this early version, some beginner mistakes were made:

- Some words, including inflections, were manually added to the anglicism list because they were present in the anglicism test set (but missing from the Wiktionary and VDS anglicism lists).
- Some words were manually filtered out of the anglicism list (e.g. words in the Wiktionary anglicism index category "0–9 – Symbole / Zeichen") since only written out words are used in the language model of the ASR system.
- In addition to the German entries and inflections, some English words and inflections were mistakenly crawled as well.

Table 4.2 shows the differences in words and pronunciations between the two versions. Compared to the base version, the initial anglicism list differs from the specific sources due to manual modification and additional English entries. All pronunciations, however, were also exclusively taken from the German Wiktionary website. Therefore, the early version will be treated as a variation of the Wiktionary approach which is referred to as wiki-v1. The wiki-v1 dictionary contained a total of 6,292 entries.

Table 4.3 on the next page shows the WERs of wiki-base and the new variation wiki-v1. Compared to wiki-base, wiki-v1 shows slightly better results for the test sets "Anglicisms 2020" (Δ 0.01 %) and "Challenging Broadcast 2018" (Δ 0.03 %). Table A3 on page 115, Table A4 on page 116 and Table A5 on page 117 show the detailed comparison of the results for the specific audio files.

One reason for differences in the hypotheses was that certain entries were exclusive to the pronunciation dictionary of one respective model. Sample 4.4 on the next page shows the benchmark results of segment 13 in "Rezo - Die Zerstörung der CDU". The word

| Model | Anglicisms 2020 (%) | German Broadcast 2020 (%) | Challenging Broadcast 2018 (%) |
|---|---|---|---|
| baseline | 15.80 | **6.56** | 10.84 |
| wiki-base | 15.72 | 6.57 | 10.83 |
| wiki-v1 | **15.71** | 6.57 | **10.80** |

**Table 4.3:** WERs for the baseline, wiki-base and wiki-v1 ASR models.

**Sample 4.4:** Segment 13 in "Rezo - Die Zerstörung der CDU" from test set "Anglicisms 2020".

| | | | | | |
|---|---|---|---|---|---|
| **reference** | und | noch'n | kleiner | Funfact | am Rande |
| baseline | und | noch ein | kleiner | | am Rande |
| wiki-base | und | noch ein | kleiner | | am Rande |
| wiki-v1 | und | noch ein | kleiner | Funfact | am Rande |

"Funfact" was only correctly recognized in the wiki-v1 hypothesis. "Funfact" is contained in the baseline dictionary with the pronunciation `/f U n f a k t/`. It is not contained in the wiki-base dictionary. The wiki-v1 dictionary contains the word with the pronunciation `/f a n f E k t/` which is different to the entry in the baseline dictionary. As the latter pronunciation matches the way the speaker pronounced the word "Funfact", it could correctly be recognized by the wiki-v1 model.

Another reason for differences in the hypotheses were changes within the WFST of the ASR model. Sample 4.5 on the following page shows the benchmark results of segment 39 in "tagesschau 20:00 Uhr, 03.02.2020". Here, the word "ist" was correctly recognized in the baseline model and wiki-base, but it was substituted with the word "in" in wiki-v1. Interestingly, the wiki-base anglicism dictionary contains the additional pronunciation `/Q I n/` for "in", but this entry is also contained in the baseline dictionary and the doubled-entry did not seem to have an effect on the WFST path in this case. However, the word "Iowa" is contained with two different pronunciations in the wiki-v1 anglicism dictionary. In this case, not the word in question, but the word following it caused changes within the WFST.

### 4.3.2 Variation 2: Combination of Both Dictionaries

Due to the observed differences in performance, the anglicism dictionaries of wiki-base and wiki-v1 were merged to build the new variation wiki-v2. The entries that were unique to wiki-base and wiki-v1 (see Table 4.2 on the previous page) are now combined in the new supplementary dictionary of wiki-v2. The wiki-v2 dictionary contained a total of

**Sample 4.5:** Segment 39 in "tagesschau 20:00 Uhr, 03.02.2020" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | [...] Marathonlauf der Vorwahlen ist Iowa die erste Etappe |
| baseline | [...] Marathonlauf der Vorwahlen ist Iowa die erste Etappe |
| wiki-base | [...] Marathonlauf der Vorwahlen ist Iowa die erste Etappe |
| wiki-v1 | [...] Marathonlauf der Vorwahlen in Iowa die erste Etappe |
| wiki-v2 | [...] Marathonlauf der Vorwahlen in Iowa die erste Etappe |

| Model | Anglicisms 2020 (%) | German Broadcast 2020 (%) | Challenging Broadcast 2018 (%) |
|---|---|---|---|
| baseline | 15.80 | 6.56 | 10.84 |
| wiki-base | 15.72 | 6.57 | 10.83 |
| wiki-v1 | 15.71 | 6.57 | **10.80** |
| wiki-v2 | **15.70** | **6.51** | 10.83 |

**Table 4.4:** WERs for the baseline, wiki-base, wiki-v1 and wiki-v2 ASR models.

9,802 entries.

Table 4.4 shows the WERs of the baseline and all models from the Wiktionary approach. It was expected that the respective best result of either model would be achieved in the merged model, but interestingly, this was not always the case. The detailed benchmark results of all test sets can be found in Table A3 on page 115, Table A4 on page 116 and Table A5 on page 117. In "Rezo - Zerstörung der CDU", "tagesthemen 22:15 Uhr, 18.02.2020" and "Venix - Tech News 94", the respective best result was achieved in wiki-v2. The new model performed worse for "tagesschau 20:00 Uhr, 03.02.2020" than wiki-base and showed the same (lower) results as wiki-v1, but this was caused by the different WFST due to the additional entries for "Iowa" which was explained in the previous subsection.

An example for positive differences caused by changes within the WFST can be seen in Sample 4.6 on the following page which shows the benchmark results of segment 25 in "tagesthemen 22:15 Uhr, 18.02.2020" from test set "Anglicisms 2020". In this example, an English phrase is cited in a news segment about the American elections. Even though word "I" does not seem like a typical anglicism, the VDS Anglizismenindex considers it as one because it is contained in phrases like "I like" and "I love it" (Elfers, 2020). The baseline dictionary as well as the wiki-v1 anglicism dictionary did not contain this anglicism, so the German homophone "Ei" was recognized instead. Both "I" and "Ei" are listed with the exact same pronunciation `/Q aI/` in the respective dictionary. Because

**Sample 4.6:** Segment 25 in "tagesthemen 22:15 Uhr, 18.02.2020" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | sein Wahlspruch I get it done ich schaffe das |
| baseline | sein Wahlspruch Ei werde dann ich schaffe das |
| wiki-base | sein Wahlspruch I werde dann ich schaffe das |
| wiki-v1 | sein Wahlspruch Ei werde dann ich schaffe das |
| wiki-v2 | sein Wahlspruch I werde dann ich schaffe das |

**Sample 4.7:** Segment 13 in "Venix - TechNews 94" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | [...] fünf neue Smartwatches |
| baseline | [...] fünf neue Smart Watches |
| wiki-base | [...] fünf neue Smart Watches |
| wiki-v1 | [...] fünf neue Smartwatches |
| wiki-v2 | [...] fünf neue Smartwatches |

"I" was contained in the dictionary in case for wiki-base and wiki-v2 and was correctly recognized in this example, it shows that the WFST of the language model put a stronger weight on its possibility than it did for "Ei", making it a better candidate for the word sequence. Hence, the additional word mapping of the pronunciation `/Q aI/` lead to a better recognition result.

In most cases, the additional entries that used to be unique to the respective models caused improvements in the recognition results of the new model. Sample 4.7 shows segment 13 in "Venix - TechNews 94" from test set "Anglicisms 2020". The word "Smartwatch" is neither contained in the baseline dictionary, nor in the wiki-base supplementary dictionary. It is, however, contained in the supplementary dictionary of wiki-v1. Due to the merge, it was also added to the supplementary dictionary of wiki-v2. In this example, the incorrect hypotheses do not seem as grave because the two word components "Smart" and "Watch" of the compound "Smartwatch" were recognized correctly since they were contained in the baseline dictionary separately. The example still proves that this additional entry in the pronunciation dictionary caused improvements for the recognition result.

### 4.3.3 Anglicism Recognition Results

Table 4.5 on the following page shows the number of recognized anglicism entities and the calculated EER of all models created with the Wiktionary approach by applying the test set "Anglicisms 2020". The models wiki-v1 and wiki v-2 recognized the most anglicisms ($\Delta$ 12) and hence show the best EER. Since wiki-v2 was created by merging

| Model | Recognized Entities | EER (%) |
|---|---|---|
| baseline | 824 | 39.50 |
| wiki-base | 834 | 38.77 |
| wiki-v1 | **836** | **38.62** |
| wiki-v2 | **836** | **38.62** |

**Table 4.5:** EERs based on a total of 1,362 anglicism entities for the baseline, wiki-base, wiki-v1 and wiki-v2 models after applying the test set "Anglicisms 2020".

the dictionaries of wiki-base and wiki-v1, the two additionally recognized anglicisms compared to wiki-base must have exclusively been contained in the wiki-v1 dictionary. The EER results for the specific audio files are shown in Table A6 on page 118.

Looking at the WER values, wiki-v2 generally performed best even though wiki-v1 recognized the same amount of anglicisms. The difference is small ($\Delta$ 0.01 %), but it shows that the recognized anglicisms did not solely contribute to the WER score. The different WFSTs influenced by all entries in the pronunciation dictionary contributed to the WER score as well.

Chapter 5

# Approach 2: Comparing German and English G2P Results

**Contents**


5.1 Creating a P2P model . . . . . 52
5.2 Implementation . . . . . 61
5.3 Evaluation . . . . . 63
5.4 Variation: Detecting Anglicisms Based on Crawl Results . . . . . 67
5.5 Anglicism Recognition Results . . . . . 71


A monolingual German G2P model is trained on a German pronunciation dictionary that may contain a small amount of loanwords, but mainly pure German words. The conversion rules are hence based on the German language and hence reflect German pronunciations of grapheme sequences. When converting a foreign word with this model, the resulting pronunciation may not reflect the actual, canonical pronunciation. The more the source languages conversion rules differ from the German ones, the less accurate the phoneme sequence generated by a monolingual German G2P should be. The accuracy should be reflected in the confidence measure of a pronunciation result: the lower the value, the less sure the model is about the resulting pronunciation to be correct. A low confidence value could be used as an indicator for foreign words. Since anglicisms are words of English heritage, this lead can be used to generate a more accurate pronunciation. A monolingual English G2P can be used to generate an additional pronunciation for each word. Compared by their confidence measure, the best pronunciation can be chosen. To prevent English pronunciations to win over German pronunciations in a word list that mainly contains pure German words, a threshold can be implemented to favor the German pronunciation.

In this approach, G2P results of both an English and a German G2P will be compared to each other. Based on the anglicism list, both an English and a German pronunciation will be generated. The respective pronunciation with the highest confidence measure will be chosen to compile a supplementary anglicism dictionary. Additionally, a word list derived from crawl results by Fraunhofer IAIS will be used for a variation of this approach. An additional confidence measure threshold will be used to generally favor the German results over the English ones since the list is complied based on German websites. The

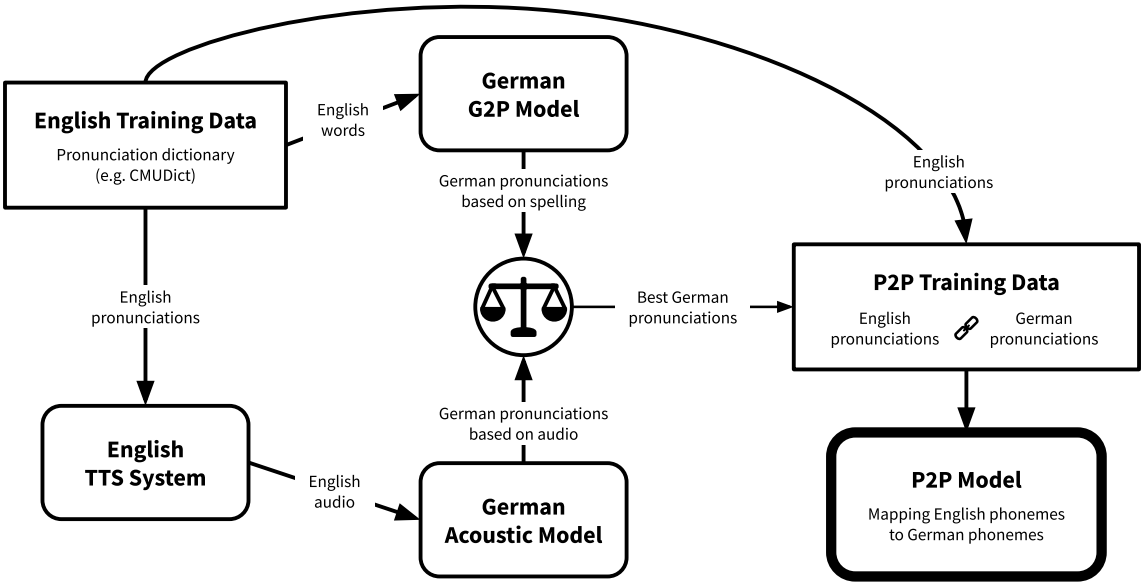


**Figure 5.1:** Process of creating a P2P model that maps English to German pronunciations. Based on an English pronunciation data, German pronunciation equivalents are created by both a G2P and an acoustic model. The best German pronunciation is chosen to be paired with the respective English pronunciation. The pronunciation pair is added to the P2P training data which is used to create the P2P model.

words that have a German pronunciation with a confidence measure below the threshold will be extracted. For those words, both an English and a German pronunciation will be generated and compared against each other by their confidence measure. If the English result is best, the word and pronunciation will be chosen to compile an English subset of the generated pronunciation dictionary that will be used as a supplementary dictionary. The English pronunciations of both the anglicism and the dictionary word list results will be mapped to German phonemes to comply with the ASR system.

## 5.1 Creating a P2P model

In this approach, English pronunciations are used that potentially contain phonemes that are not present in the German language. The acoustic model in the ASR system at Fraunhofer IAIS, however, only handles phonemes from the German phoneme set. Therefore, a system has to be created that transforms English phonemes to their nearest German counterpart.

Based on the findings and experiments of Patel et al. (2018) and Bruguier et al. (2017), training data has been prepared to create a Phoneme-to-Phoneme (P2P) model[1] that is able to map English phonemes to their German equivalent. Figure 5.1 describes the

[1]A P2P model is essentially a G2P model that takes a phoneme sequence as input instead of a grapheme sequence. Hence, it can be build to map target language phoneme sequences to source language phoneme sequences.

process of creating the P2P model. English and German training data is needed to produce a phoneme lexicon that contains English pronunciations and their respective German equivalents. Based on this mapping dictionary, a P2P model is trained which will then be able to convert English phonemes in ARPABET notation to German phonemes in BAS-SAMPA notation.

### 5.1.1 Data Collection

As seen in Figure 5.1 on the previous page, the following data was needed to prepare training data for the P2P model:

- **English training data:** An English pronunciation dictionary (e.g. CMUdict) that contains words and their respective canonical pronunciation
- **TTS-synthesized English training data:** Synthesized pronunciations of the English training data as audio files
- **German training data based on spelling:** German pronunciations based on the words of the English training data, generated with a German G2P model
- **German training data based on audio:** German pronunciations based on the TTS audio of the English training data, generated with a German phoneme recognizer that is based on the acoustic model

#### English Training Data

An English data source is needed that provides grapheme and phoneme sequence combinations. According to Patel et al. (2018), the source should be large and diverse to make sure it provides all pronunciation combinations and variations that can occur in the language. The CMUdict complies with those requirements and will hence be used as English training data. The CMUdict version 0.7b was used which contains 133,779 entries. The entries were cleaned from notations that were not necessary for training. Listing 5.1 on the following page shows some example entries before and after cleaning. In ARPABET notation, stress markers are added as a suffix to vowels in form of a number between 0 and 2. The suffixes stand for *no stress* (0), *primary stress* (1) and *secondary stress* (2). Those stress markers have been removed from the vowel phonemes since they "are usually omitted in ASR acoustic modelling" (Milde et al., 2017). Also, words that have multiple pronunciation variants are listed with a variation number in brackets behind the word. Those variation brackets were also removed as they would be read as additional graphemes by the G2P.

**Listing 5.1:** Original CMUdict pronunciations and cleaned versions without stress markers and variant brackets.

```
# Original:
ASPIRANTS       AE1 S P ER0 AH0 N T S
ASPIRANTS(1)    AH0 S P AY1 R AH0 N T S
ASPIRANTS(2)    AE1 S P ER0 AH0 N S
ASPIRANTS(3)    AH0 S P AY1 R AH0 N S

# Cleaned:
ASPIRANTS   AE S P ER AH N T S
ASPIRANTS   AH S P AY R AH N T S
ASPIRANTS   AE S P ER AH N S
ASPIRANTS   AH S P AY R AH N S
```

#### TTS-synthesized English Training Data

The pronunciations from the English training data have to be synthesized to create audio data as input for the German acoustic model. Instead of having a human read and record the pronunciations, Patel et al. (2018) suggests generating audio with a text-to-speech system since it will ensure the exact phoneme sequence is synthesized. Based on a PHP script[1] by Santos (2015), the ARPABET pronunciations from the English training data were converted to IPA notation so they could be processed by a TTS system. Amazon Polly[2] and Microsoft Azure Text-to-Speech[3] were used to synthesize the CMUdict pronunciations. They were generated by two male (Polly Joey and Polly Matthew) and two female (Polly Salli and Azure Aria) voices, resulting in 535,116 audio files.

#### German Training Data Based on Spelling

To generate the German phoneme equivalents of the CMUdict grapheme sequences, both the spelling and the pronunciation should be considered (Patel et al., 2018). To get the German pronunciation based on the spelling of a word, the German Sequitur G2P model trained by Fraunhofer IAIS was used. A list was created that only contained the grapheme sequences from the English training data. The CMUdict grapheme sequences are written in uppercase which resulted in the G2P model spelling out the letters in the resulting pronunciation (see Listing 5.2 on the next page). To avoid this behavior, the casing was changed to title case using the `capwords` method from the python library `string`[4] that

[1] https://github.com/wwesantos/arpabet-to-ipa/blob/043a5050d43724194a5734037397279577cddef7/src/App.php

[2] https://aws.amazon.com/de/polly/

[3] https://azure.microsoft.com/de-de/services/cognitive-services/text-to-speech

[4] https://docs.python.org/3/library/string.html

handles the casing of apostrophes correctly (e.g. “Julia’s” instead of “Julia’S”). Based on the list with updated casings, the German pronunciations including their respective confidence values were generated.

#### German Training Data Based on Audio

A live recognizer application that utilizes the acoustic model from the ASR system at Fraunhofer IAIS was used to get the German pronunciation based on the TTS-generated audio files. The live recognizer recognizes and produces a text string based on an audio speech utterance. For this work, a colleague implemented a feature that allows the live recognizer to return phoneme output instead. The feature was still in beta testing and hence not fully developed when it was utilized in this work, so quality loss was expected for the phoneme sequence results. The TTS-generated audio files were sent to the live recognizer client by using a Python script that was slightly modified to return the desired information in the JSON output. Listing 5.3 shows an example JSON result returned by the client which i.a. contains the recognized phoneme string and its probability value. The result phoneme string is formatted in BAS-SAMPA with an additional notation following the phoneme separated by an underscore that declares the phonemes position in the sequence: `_B` for beginning, `_I` for intermediate and `_E` for end.

The results were separated by the respective voice used to produce the audio file. Unfortunately, some words have not been processed by the live recognizer:

- Polly Joey: 170 (0.12 %)
- Polly Matthew: 170 (0.12 %)

**Listing 5.2:** Different German Sequitur G2P outputs for words in uppercase and title case.

```
# Uppercase
COW     ts e: Q o: v e:
OTTER   Q o: t e: t e: Q e: Q E6
WHALE   v e: h a: Q a: Q E l Q e:

# Title case
Cow     k aU
Otter   Q O t 6
Whale   v a: l @
```

**Listing 5.3:** JSON output by the live recognizer.

```
{"likelihood": 164.359, "confidence": 0.9903229676489286, "transcript": "v_B e:_I l_E.", "
    ↪ word": "WHALE"}
```

- Polly Salli: 224 (0.16 %)
- Azure Aria: 165 (0.12 %)

However, since a total of 133,779 words were available for each voice, the issue only affected a very small fraction of the total results. Also, the words from the respective missing results have all been successfully processed for at least two other voices. Therefore, the missing results have been disregarded to save time for other important tasks.

### 5.1.2 Implementation

The Sequitur G2P framework was used to build a P2P model. It offers a P2P option (`--phoneme-to-phoneme`) that enables the training of a P2P instead of a G2P model. This option is not documented, so the usage had to be found out by testing and consulting the source code. While the default G2P option only needs one respective file containing the train, validation and test data, two files are needed for the P2P option. The first data file should contain grapheme sequences and the source pronunciation (English ARPABET) and the second data file should contain grapheme sequences and the target pronunciation (German BAS-SAMPA). In the bash command, the path to both files are declared separated by a colon. An example command for training the first iteration of a P2P model is shown in Listing 5.4.

To produce the final training data for the P2P model, a python script was written that processes the data created in Section 5.1.1 on page 53. This includes the data generated with the German Sequitur G2P model (G2P pronunciations) and the data generated by the live recognizer (AM pronunciations) which resulted in five data files. First, the AM pronunciations of all four voices were compared against each other to find the best matching pronunciation. The Levenshtein distance compared to the original English pronunciation was used as the main comparison metric. The German and the English phoneme inventory have 80 % phonemes in common (Patel et al., 2018), so the result with the most matching phonemes compared to the original pronunciation is expected to be the best German equivalent. For this comparison, all pronunciations were converted to IPA notation since they existed in two different notations (ARPABET and BAS-SAMPA). The pronunciation with the lowest Levenshtein distance was chosen as the best pronunciation from the AM pronunciations. If multiple pronunciations had the same low Levenshtein

**Listing 5.4:** Bash command to train the first iteration of a P2P model with Sequitur G2P.

```
g2p.py --phoneme-to-phoneme --train train_en.dict:train_de.dict --devel val_en.dict:val_de
    ↪ .dict --write-model model-1
```

distance, the pronunciation with the highest confidence value was chosen.

To get the final German pronunciation, the best pronunciation AM pronunciation was compared against the G2P pronunciation for each word by comparing their confidence values. Additionally, a weight parameter was used to be able to weight the results differently. With this parameter, it was able to favor the AM pronunciation over the G2P pronunciation and vice versa. Listing 5.5 shows the selection of the final pronunciation in pseudocode.

Having the final German pronunciation equivalents for the English pronunciations, the training and validation data for the P2P model creation could be compiled. 95 % of the total training data was used as train set and the remaining 5 % were used as validation set. For each data set, two files were created containing (1) the grapheme sequence with original English pronunciation in ARPABET notation and (2) the grapheme sequence with generated German pronunciation in BAS-SAMPA notation.

### 5.1.3 Evaluation

Nine weight values have been chosen by the binary search method depending on their test results. For each weight, the P2P models have been trained for 6 iterations. The models were saved after each iteration. All intermediate models were included in the evaluation process to avoid having an iteration that does not perform as well as the previous one.

To evaluate the resulting P2P models for their future performance in a realistic application, data was needed that both had an English pronunciation and its German counterpart. Wiktionary provides anglicism pronunciations that have already been crawled for the Wiktionary approach (see Chapter 4 on page 42), so the best performing dictionary from model wiki-v2 has been used to provide the German pronunciations for the P2P test set. The list consisted of 9,802 entries. The respective English pronunciations were taken from CMUdict. After matching the entries, the resulting test set consisted of 2,267 pronunciation pairs.

Table 5.1 on the next page shows the PER values for the respective best model after applying the Wiktionary test set. AM weight 0 implies that only the G2P pronunciations were used; AM weight 1 implies that only the AM pronunciations were used. Overall, AM

**Listing 5.5:** Selection of the final German pronunciation equivalent in pseudocode.

```
am_weight = 0.25
if (am_weight * lr_pronunciation) < ((1 - am_weight) * g2p_pronunciation):
    choose lr_pronunciation
else:
    choose g2p_pronunciation
```

| AM Weight | Iteration of Best Model | PER (%) |
|---|---|---|
| 0 | 5 | **19.65** |
| 0.1 | 5 | 21.38 |
| 0.2 | 4 | 21.68 |
| 0.25 | 4 | 21.83 |
| 0.3 | 4 | 22.31 |
| 0.4 | 5 | 23.49 |
| 0.5 | 3 | 27.98 |
| 0.6 | 2 | 31.02 |
| 0.75 | 1 | 34.27 |
| 1 | 2 | 36.02 |

**Table 5.1:** Best PER results for each the P2P models of each AM weight value after applying the Wiktionary test set.

| Iteration | PER (%) |
|---|---|
| 1 | 43.45 |
| 2 | 34.82 |
| 3 | 33.73 |
| 4 | 31.20 |
| 5 | 29.87 |
| 6 | 29.63 |

**Table 5.2:** PER values of the P2P models of each iteration for AM weight 0.5 after applying the validation set.

weight 0 performed best with a PER of 19.65 %. While 6 iterations were trained for each weight, all AM weights reached their best model at a lower iteration. It was noticeable that except for AM weight 0.4, the number of the respective best iteration decreased with increasing AM weight. This means that the model got worse with each following iteration, implying that there might be issues with the data quality.

Tested on the validation set, the PER values actually improved with the number of iterations for all AM weights. Table 5.2 shows the PER values for AM weight 0.5 exemplarily after applying the validation set. In contrast to the Wiktionary test set which comes from a different source, the validation set contains similar data as the train set since they come from the same source. The improving PER values per iteration were observed for all AM weights and it shows that the models are in fact steadily improving on the data.

| AM Weight | Min. PER (%) | Max. PER (%) | Mean PER (%) | $\sigma$ PER (%) |
|---|---|---|---|---|
| 0 | 14.34 | 14.78 | 14.56 | 0.150280 |
| 0.25 | 21.66 | 21.95 | 21.80 | 0.097365 |
| 0.5 | 30.25 | 30.71 | 30.46 | 0.169658 |
| 0.75 | 31.26 | 31.54 | 31.42 | 0.107070 |
| 1 | 30.30 | 30.55 | 30.42 | 0.099720 |

**Table 5.3:** K-fold cross validation results for the models of chosen AM weights.

#### K-Fold Cross Validation

To further evaluate the quality of the model approach and data sets, k-fold cross validation has been performed for AM weights 0, 0.25, 0.5, 0.75 and 1 exemplarily. As the data is the only variable that is different for these models, the results should show if the training data itself is consistent for each respective AM weight. A $k$ value of 5 has been chosen as this value has "been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance" (James et al., 2013, p.186). Hence, the data has been shuffled and separated into five folds. For each AM weight, five models have been trained that each used a dedicated fold as test set and all respective other folds as train set. All models have been trained for six iterations.

Table 5.3 shows the results of the k-fold cross validation for each chosen AM weight. Even the minimum and maximum PERs per model are quite similar, hence showing a low standard deviation for all models. This low spread among the respective PER results implies that the model approach and the data are robust.

Looking at the mean PER values, it is noticeable that the values increase with increasing AM weight, except for the value for AM weight 1. While the data for AM weight 0 and 1 only comes from one data source respectively, data for AM weights 0.52, 0.5 and 0.75 consist of both the G2P and the AM pronunciations. AM weight 0 having a better mean PER than am weight 1 implies that the data for AM weight 0 is more consistent because the value is much lower even though it was trained for the same amount of iterations. The model seems to improve slower on the data consisting of the AM pronunciations than it does on the data consisting of the G2P pronunciations. The mean PER of the AM weight 0.25, 0.5 and 0.75 models hence seem to increase the more AM pronunciations the data consists of.

#### Interpreting the PER

While the PER results give an idea on how well the models perform, they are based on matching the exact phoneme sequences. As mentioned in Section 3.3.4 on page 41, some

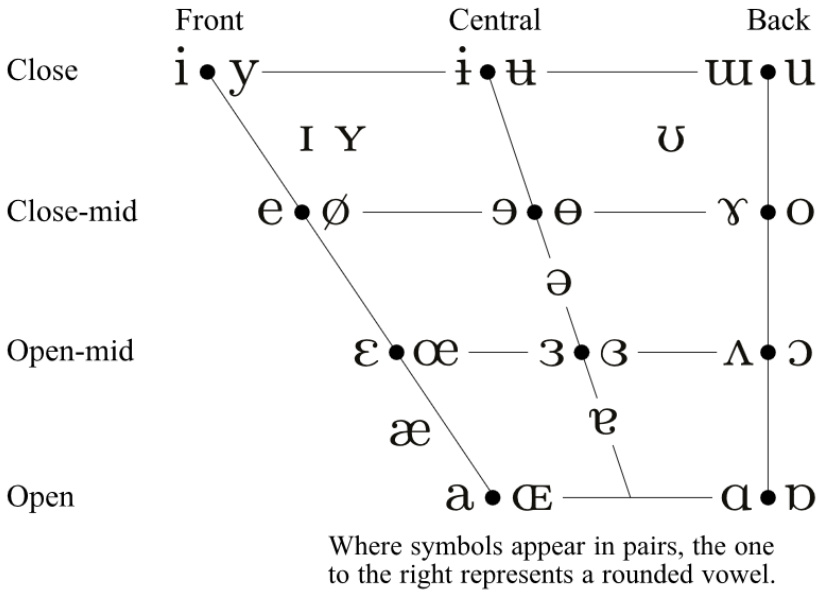


**Figure 5.2:** Vowel diagram from the International Phonetic Alphabet (International Phonetic Association, 2015)

phonemes are more similar to each other than others, making the actual errors less grave. This can be most simply explained with the example of vowels. The vowel diagram in Figure 5.2 illustrates the closeness (y-axis) and backness (x-axis) needed to form a vowel in the vocal tract. In case of pairs, the symbol on the right represents a rounded vowel which means that it is formed with rounded lips. Comparing the phone /ɜ/ to /ə/, they both lie very close to each other in the diagram, being formed central, mid to open-mid and unrounded. Comparing /ɜ/ to /u/, however, shows that they are much farther apart form each other with /u/ being formed in the back, closed and rounded.[1]

The entries in the pronunciation dictionary of an ASR system represent the canonical pronunciation of a word plus a few pronunciation variations in some cases. It is possible that substituting a phoneme with a similar one leads to a pronunciation variation rather than a wrong pronunciation. Therefore, some phoneme sequences that contained phoneme errors when comparing them to a list of canonical pronunciations like in the Wiktionary test set might in fact be realistic pronunciation variations. This might especially be the case for the pronunciations generated by the live recognizer as it captured what was heard in the audio file. To inspect this assumption, five pronunciation dictionaries will be created, each generated with a P2P model that is trained with data of a different AM weight. Each dictionary will be included in a dedicated ASR model, making it possible to evaluate the different AM weights of the P2P training data under realistic conditions. The AM weights 0, 0.25, 0.5, 0.75 and 1 will be used.

[1]This website provides an interactive IPA chart that plays sound when clicking on the respective IPA symbol: `https://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/`

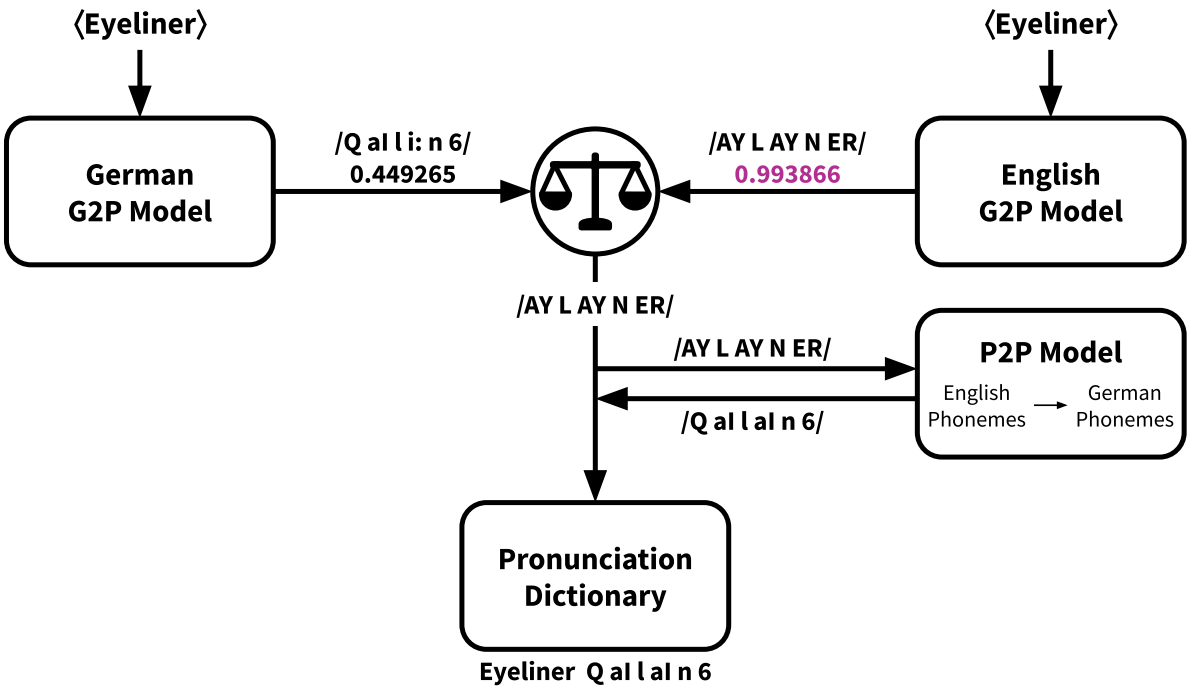


**Figure 5.3:** Example of the comparative process used on the anglicism "Eyeliner". The grapheme sequence is put into both the German and the English G2P models. The respective output phoneme sequences are then compared to each other by their confidence measure. If the English pronunciation wins, it is mapped to German phonemes by the P2P model. The resulting pronunciation is added to the pronunciation dictionary.

## 5.2 Implementation

After the five P2P models were created, the actual implementation of the comparative approach could begin. Figure 5.3 shows the desired process based on the example anglicism "Eyeliner". First, an English G2P model was needed to produce English pronunciation results for the anglicism list. Since Fraunhofer IAIS uses a German Sequitur G2P for generating pronunciations, an English Sequitur G2P model was trained to reliably compare their confidence measures against each other. The most recent version of CMUdict[1] with 135.154 total entries was used for training with a 80/20 split for the train and test data sets. 5 % of the train set was held back as a validation set.

Before the split, CMUdict was cleaned from variation brackets after the grapheme sequences, stresses and comments (see Section 5.1.1 on page 53). All entries that represent different pronunciation variants of a word were included in the training set as they give valuable information for differing phoneme mappings for the same grapheme sequences and they would interfere with testing since only one pronunciation is generated to evaluate the error rates. Also, the casing was changed to lowercase as all grapheme sequences

[1] https://github.com/cmusphinx/cmudict/blob/bfb3b05747125a2405a79f1afd73a42a90ba8c3a/cmudict.dict

| Model | PER (%) | WER (%) |
|---|---|---|
| model-1 | 43.10 | 97.37 |
| model-2 | 18.13 | 64.97 |
| model-3 | 10.79 | 42.82 |
| model-4 | 7.94 | 32.33 |
| model-5 | 7.24 | 29.55 |
| model-6 | 6.94 | 28.52 |
| **model-7** | **6.87** | **28.27** |
| model-8 | **6.87** | 28.28 |
| model-9 | **6.87** | 28.30 |
| model-10 | **6.87** | 28.31 |

**Table 5.4:** PER and WER results of each iteration for the English Sequitur G2P model after applying the test set.

were formatted as uppercase in CMUdict.

The English G2P model was trained for 10 iterations (see Table 5.4). The model created after iteration 7 (model-7) was chosen as it showed the best PER (6.87 %) and WER (28.27 %) values. The words in the anglicism list were split in case they contained a whitespace as the resulting pronunciation dictionary only lists single words. Words containing letters that are not present in the English alphabet (ä, ö, ü, ß) were filtered out since the English G2P does not recognize those characters which reduced the list from 18,967 to 17,035 words. Also, since the English G2P model was trained on lowercase words, the words from the anglicism list were formatted to lowercase as well. Based on this list, the English pronunciations of the anglicism list including their confidence measures were produced with model-7.

Next, the German pronunciations of the anglicism list had to be created. Fraunhofer IAIS already has a well-performing German Sequitur G2P model which is used for creating the pronunciation dictionary for the ASR system. Hence, this model was used to produce the German pronunciations including their confidence measures based on the split anglicism list containing 18,967 words.

Unfortunately, the anglicism list contained some words that at least one of the G2P models was not able to handle. Both models had problems with punctuation marks (e.g. “.”, “,”, “/”). The German model was not able to process a few words containing numbers (e.g. “360-Grad-Panaroma”, “MP3-Format”, “Ü-30-Party”) even though it was able to generate pronunciations for words like “20-Cent-Münze” and “Top-10”. In total, the English G2P model was not able to process 36 of 17,035 words (0.21 %) and the German

G2P model was not able to process 50 of 18,967 words (0.23 %).

After the results for both G2P models were generated, the final pronunciation dictionary could be created. Since the German G2P is trained on the PHONOLEX core data set, it was additionally consulted since it contains manually checked results by linguist professionals. 1,435 words from the split anglicism list were contained in PHONOLEX core. Of those entries, 75 pronunciations differed from the German G2P results which were hence chosen as the first entries in the final anglicism list.

Afterwards, the English and German G2P results were compared against each other by their confidence measure. The result with the highest value was chosen for the final pronunciation dictionary. For words that were not available in the respective other result list (e.g. filtered words with non-English letters), the pronunciation was directly chosen as no comparison was possible. The resulting dictionary contained 18.917 entries of which 10,216 pronunciations were English, 8,626 were German and 75 were directly taken from PHONOLEX.

As the English pronunciations were still written in ARPABET notation, they had to be mapped to their German expressions. As mentioned in Section 5.1.3 on page 57, five AM weights were chosen to be tested in the ASR model, so the English pronunciations were mapped to BAS-SAMPA with a P2P model that was trained with each of those AM weights, resulting in 5 different result sets. Added to the final entries that were already consisting of German pronunciations, five final pronunciation dictionaries were created.

## 5.3 Evaluation

Based on the created anglicism dictionaries, five ASR models were created. The dictionaries contained 18.917 entries[1]. Table A7 on page 119 shows 20 example entries from the respective anglicism dictionaries. While for comp-0 the training data for the respective P2P model was generated exclusively from the G2P pronunciations, the training data for the P2P model used in comp-1 was exclusively trained with the AM pronunciations. The P2P models of comp-0.25, comp-0.5 and comp-0.75 were trained by using mixed data of the G2P and AM pronunciations, so the corresponding dictionaries showed varying results depending on which component weighted more for creating the training data of the respective P2P model. The pronunciations for the word “Crowdfundigs” best display the weight influence of the P2P training data as they all differ from each other (see Table 5.5 on the next page).

For comp-0, it is noticeable that the training data for the used P2P model was exclusively

[1]The dictionaries for models comp-0.75 and comp-1 were missing the entry for the word “ear” because the P2P model was not able to return a result for the ARPABET pronunciation `/IH R/`. Hence, the dictionaries for models comp-0.75 and comp-1 only contain 18.916 entries.

| Model | Pronunciation |
|---|---|
| comp-0 | k r u: t f U n d I N s |
| comp-0.25 | k r a U t f U n d I N s |
| comp-0.5 | k r o U t f U n d I N s |
| comp-0.75 | k 96 f a n d E N s |
| comp-1 | k r @ U t f a n d E N s |

**Table 5.5:** Pronunciations for the anglicism "Crowdfundings" generated by the comparative ASR models.

| Model | Anglicisms 2020 (%) | German Broadcast 2020 (%) | Challenging Broadcast 2018 (%) |
|---|---|---|---|
| baseline | 15.80 | **6.56** | 10.84 |
| comp-0 | 15.80 | 6.57 | 10.86 |
| comp-0.25 | 15.67 | 6.57 | **10.82** |
| comp-0.5 | **15.59** | 6.57 | **10.82** |
| comp-0.75 | 15.67 | 6.57 | 10.83 |
| comp-1 | 15.67 | 6.57 | 10.83 |

**Table 5.6:** WERs for the baseline and the comparative ASR models.

generated by the German G2P model. Pronunciations like `/j o: k @ s/` ("Jokes") and `/S i: n h e a: ts/` ("Skinheads") are some typical examples for what it looks like when German pronunciation rules are applied to English words. For comp-1, the influence of the English TTS audio was observed. The pronunciations `/Q E6 p k E6 S I n/` ("abcashen") and `/ts v E d 2: s/` ("Sweaters") are typical examples for how an English speaker would pronounce the respective words. In the case of "abcashen", only the word stem "cash" is an English word while the German prefix "ab" and the German suffix "en" transform the stem to a verb used in the German language. Hence, "ab" and "en" were pronounced by using English pronunciation rules by the English TTS voice that was used to generate the data. Unfortunately, some pronunciations like `/z k I6 n h E l s/` ("Skinheads"), `/k u a S s/` ("Squashs") and `/v e: a e: s/` ("VIPs") show that the P2P training data might have had some quality issues. The pronunciation for "Skinheads" sounds like it was generated from the word "Skinhells" instead since `⟨d⟩` was mapped to the phoneme `/l/`. For "Squashs" and "VIPs", phonemes are missing, making the pronunciations rather sound like "Quashs" and "VAEs". However, since the same acoustic model used for generating the training data for the respective P2P model was also used by the ASR model, it is possible that those pronunciations might lead to correct anglicism recognitions regardless.

**Sample 5.1:** Segment 9 in "Venix - TechNews 94" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | ziemlich nices Feature eigentlich [...] |
| baseline | ziemlich neues Feature eigentlich [...] |
| comp-0 | ziemlich neues Feature eigentlich [...] |
| comp-0.25 | ziemlich nices Feature eigentlich [...] |
| comp-0.5 | ziemlich nices Feature eigentlich [...] |
| comp-0.75 | ziemlich nices Feature eigentlich [...] |
| comp-1 | ziemlich nices Feature eigentlich [...] |

Table 5.6 on the preceding page shows the WERs of the five comparative ASR models that contain the aforementioned anglicism dictionaries compared to the baseline model. For test set "Anglicisms 2020", all models except for comp-0 showed improvements with comp-0.5 being the best one ($\Delta$ 0.19 %). For test set "German Broadcast 2020", all comparative models slightly decreased in WER by 0.01 percentage points. The benchmarking results for test set "Challenging Broadcast 2018" showed mixed results. While model comp-0 slightly decreased by 0.02 percentage points, the other four models showed slightly increased WER values with models comp-0.25 and comp-0.5 being the best ($\Delta$ 0.02 %). Table A8 on page 120, Table A9 on page 121 and Table A10 on page 122 show the detailed comparison of the results for the specific audio files. Model comp-0.5 showed the best overall performance, showing that the equal mix between the German G2P pronunciations and AM pronunciations as training data for the P2P model worked best.

The most differences occurred due to newly added pronunciations in the supplementary anglicism dictionaries. Depending on if the generated pronunciations in the anglicism dictionaries conform to how the speakers actually pronounce them, they will be recognized by the ASR model. An example is shown in Sample 5.1 which shows the benchmark results of segment 9 in "Venix - TechNews 94" from test set "Anglicisms 2020". The word "nices" was not recognized by the baseline and the comp-0 model, even though it is listed in the baseline dictionary as `/n I s @ s/` and the comp-0 anglicism dictionary as `/n I s s/`. However, both pronunciations do not reflect the actual pronunciation of this anglicism. The word is listed with the pronunciation `/n aI s @ s/` in the dictionaries of models comp-0.25, comp-0.5, comp-0.75 and comp-1 which matched correctly with the way the speaker in this segment pronounced the anglicism.

Added entries in the pronunciation dictionary not only lead to the respective word being recognized correctly, they could also affect preceding and following words. Sample 5.2 on the next page shows an example of this phenomenon. In this segment that uses colloquial speech, the phrase "fucking Black Mirror" was only recognized by model comp-0.5. The

**Sample 5.2:** Segment 19 in "Rezo - Wie Politiker momentan auf Schüler scheißen" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | das ist schon fast eine fucking Black Mirror Folge |
| baseline | das ist schon fast eine Pekingblick Müller Folge |
| comp-0 | das ist schon fast eine Pekingblick Müller Folge |
| comp-0.25 | das ist schon fast eine Pekingblick Müller Folge |
| comp-0.5 | das ist schon fast eine fucking Black Mirror Folge |
| comp-0.75 | das ist schon fast eine Pekingblick Müller Folge |
| comp-1 | das ist schon fast eine Pekingblick Müller Folge |

| Model | Pronunciation |
|---|---|
| baseline | m I r o:6 |
| | m I6 @ |
| | m I6 @ r |
| comp-0 | m I r o:6 |
| comp-0.25 | m I r o:6 |
| comp-0.5 | m i r 6 |
| comp-0.75 | m i r @ r a |
| comp-1 | m i r @ r a |

**Table 5.7:** Pronunciations for the word "Mirror" in the baseline and comparative dictionaries.

recognized pronunciations `/f a k I N/` for "fucking" and `/b l E k/` for "Black" are listed in the baseline dictionary and hence available for the baseline and all comparative ASR models. However, the pronunciation for the word "Mirror" differs in all dictionaries as seen in Table 5.7.

Only the pronunciation dictionary of model comp-0.5 contains the pronunciation `/m i r 6/` which corresponds to the actual pronunciation of the anglicism "Mirror" by the speaker. With this last word missing as an option in the other ASR models, the phrase was not recognized. The word "Müller" was recognized instead which is listed with a similar-sounding pronunciation `/m Y l 6/` in the baseline dictionary. Even though the first two words had the possibility of being recognized, the WFST chose the word "Pekingblick", listed with pronunciation `/p e: k I N b l I k/` in the baseline dictionary, over them since it made more sense looking at the following word "Müller" which influenced the path weights.

For the "German Broadcast 2020" test set, all comparative models performed slightly worse than the baseline. However, this was not caused by different dictionary entries, but

**Sample 5.3:** Segment 16 in "Der Limbecker Platz in Essen" from test set "German Broadcast 2020".

| | |
|---|---|
| **reference** | [...] den Konsumenten auch immer wieder was Neues geben [...] |
| baseline | [...] den Konsumenten noch immer wieder was Neues geben [...] |
| comp-0 | [...] den Konsumenten noch immer wieder etwas Neues geben [...] |
| comp-0.25 | [...] den Konsumenten noch immer wieder etwas Neues geben [...] |
| comp-0.5 | [...] den Konsumenten noch immer wieder etwas Neues geben [...] |
| comp-0.75 | [...] den Konsumenten noch immer wieder etwas Neues geben [...] |
| comp-1 | [...] den Konsumenten noch immer wieder etwas Neues geben [...] |

by a phenomenon that was observed in the file "Der Limbecker Platz in Essen". Sample 5.3 shows an extract of section 16 in the mentioned file. The word "was" is used as a colloquial form of "etwas" (*some*) in this phrase. Instead of recognizing the colloquial "was" like the baseline model did, all comparative models recognized the formal form "etwas" even though no /Q E t/ sound could be heard in the audio by human inspection. Also, none of the words listed in this example are contained in the supplementary anglicism dictionaries by the comparative models, so the baseline dictionary was applied for all word recognitions. This phenomenon was also observed in other recognition results, including models that are not mentioned in this work.

Recognizing this formal form might be caused by the language model. The language model is trained on large amounts of written language which usually utilizes formal word forms. Therefore, it might weight the formal word forms stronger than it weights the colloquial form. Also, the word "was" is only used in terms of "what" in formal language which has a different meaning and hence is used in different contexts than "etwas". Since this reasoning would apply to the baseline model as well, it is unknown why the phenomenon does not occur in the baseline result. While Sample 5.3 only shows an extract of segment 16 which in total contained 43 words, all other recognition differences were equal for the baseline and the comparative models. However, it is possible that words contained in this total segment might have caused differences in the WFST weights.

## 5.4 Variation: Detecting Anglicisms Based on Crawl Results

The anglicism pronunciation dictionaries that were evaluated in the previous section were built by comparing the English to the German pronunciations by their confidence measures. As the confidence measure expresses how certain the G2P model is about the resulting pronunciation, a low value could mean that the character combination of the grapheme sequence was challenging to map. If this was the case, a low confidence measure could be an indicator for identifying foreign words.

| Word | Confidence Measure | G2P Pronunciation | Wiktionary Pronunciation |
|---|---|---|---|
| Haus | 0.984176 | h aU s | h aU s |
| Kiste | 0.988470 | k I s t @ | k I s t @ |
| Lametta | 0.857330 | l a m E t a | l a m E t a |
| piesacken | 0.710845 | p i: z a k @ n | p i: z a k n |
| Feature | 0.149451 | f tS 6 | f i: tS 6 |
| Movie | 0.177397 | m o f i: | m u: v i |
| Practice | 0.330482 | p r a k tS s | p r E k t I s |
| Choke | 0.146076 | k o k @ | tS o U k |

**Table 5.8:** Typical German words (top 4 rows) and English words (bottom 4 rows) with their confidence measures and pronunciations generated by Fraunhofer IAIS' German G2P model compared to the respective pronunciations taken from Wiktionary, converted to BAS-SAMPA

Table 5.8 shows pronunciations and their confidence measures generated by Fraunhofer IAIS' German Sequitur G2P model compared to the respective pronunciations taken from Wiktionary. The first four rows show typical German words and the last four rows show English words. The G2P pronunciations of the German words have a high confidence measure and match (or almost match in case of "piesacken") the Wiktionary pronunciations[1]. The G2P pronunciations for the English words, however, show a low confidence measure and dramatic differences to the Wiktionary pronunciations, implying that the G2P model had challenges mapping those words. The difference in confidence measures indicates that foreign word detection could be possible with this method.

The assumption that the confidence measure can be used as an indicator for detecting foreign words will be tested with this variation. Since this work is about anglicism recognition, only German and English G2P models will be used for the result comparison. Adding G2P models of other languages (e.g. French) to the comparison would be possible as well, but it will not be tested within this work.

### 5.4.1 Data Collection

Every day, Fraunhofer IAIS crawls German websites for text segments to train their language model and expand the pronunciation dictionary. This makes it possible to quickly react to new words in the German language, e.g. caused by novel incidents like the Corona-pandemic or a new star entering the spotlight. For all words that occurred

[1]Please note that "Haus" and "Kiste" were contained in the train data set of the G2P model, hence their canonical pronunciations were already seen by the model in training phase.

more than three times in the crawl results, a pronunciation is generated automatically by the German Sequitur G2P model and added to the dictionary. Since the pronunciations are created by a German G2P, they are vulnerable to the challenges described in the previous paragraph. A word list containing 3,213,274 words that occurred more than three times in the crawl results was used to retrieve anglicisms that would potentially achieve better pronunciation results using an English G2P.

To filter out the candidates, a threshold of 0.4 was set for the confidence measure of the German G2P results to slightly favor the German results as the vocabulary mainly consists of German words. All phoneme sequences with a confidence measure below the threshold are saved in a candidate list with their respective German pronunciation and confidence measure. Based on the words in the candidate list, English pronunciations were generated with the English Sequitur G2P model (model-7) and saved including their confidence measures.

### 5.4.2 Implementation

The German pronunciations from the candidate list were compared to the English pronunciations by their confidence measure. If the value for the English pronunciation was higher than the German one, it was added to the final pronunciation dictionary.

The pronunciations were mapped to their German expressions with a P2P model that was trained with an AM weight of 0.5 since it performed best in the benchmarking result of the comparative anglicism dictionaries. The resulting dictionary contained 389,119 English pronunciations mapped to German BAS-SAMPA expressions representing possible anglicisms based on the preassigned threshold of 0.4 and comparison against the German G2P results.

### 5.4.3 Evaluation

A dedicated ASR model was created that includes the supplementary anglicism dictionary from the crawl variation. The dictionary contained a total of 389,119 entries and was hence the biggest supplementary dictionary of all models created for this work. Table 5.9 on the following page shows the WERs for the test sets “Anglicisms 2020”, “German Broadcast 2020” and “Challenging Broadcast 2018”. The comp-crawl model performed slightly better on both the “German Broadcast 2020” (Δ 0.01 %) and the the “Challenging Broadcast 2018” (Δ 0.01 %) test sets. However, it showed an increased WER compared to the baseline model for the test set “Anglicisms 2020”.

Similar to the other models tested so far, the majority of benchmark differences was caused by additional entries in the pronunciation dictionary. As an example, Sample 5.4 on the next page shows an extract of segment 8 in “Polizeigewalt gegen Demonstranten in

| Model | Anglicisms 2020 (%) | German Broadcast 2020 (%) | Challenging Broadcast 2018 (%) |
|---|---|---|---|
| baseline | 15.80 | 6.56 | 10.84 |
| comp-crawl | 15.89 | **6.55** | **10.83** |

**Table 5.9:** WERs for the baseline and comp-crawl ASR models.

**Sample 5.4:** Segment 8 in "Polizeigewalt gegen Demonstranten in Hongkong" from test set "German Broadcast 2020".

| | |
|---|---|
| **reference** | `[...] Kundgebung im Victoria Park im Stadtteil Causeway Bay [...]` |
| baseline | `[...] Kundgebung im Victoria Park im Stadtteil Corso eBay [...]` |
| comp-crawl | `[...] Kundgebung im Victoria Park im Stadtteil Causeway Bay [...]` |

Hongkong" from test set "German Broadcast 2020". The baseline model recognized the words "Corso eBay" while comp-crawl recognized the correct words "Causeway Bay". The word "Bay" was already contained in the baseline dictionary with the correct pronunciation `/b E I/` and there was no additional pronunciation in the comp-crawl dictionary. The word "Causeway" was included in the baseline dictionary as well with the pronunciations `/k aU z @ v e:/` and `/k O: z v e I/`, however, the model was not able to recognize the spoken word by those options. In the comp-crawl dictionary, the additional pronunciation `/k O6 z u e: I/` was contained which matches the way the speaker pronounced the word "Causeway". Since this pronunciation was not available in the baseline dictionary, the similar pronunciations `/k O6 z o/` ("Corso") and `/Q i: b E I/` ("eBay") were recognized and used for the respective transcriptions.

Some additional entries, however, also caused wrong recognition results for comp-crawl. Sample 5.5 on the following page shows an extract of segment 39 in "Venix - Tech News 95" from test set "Anglicisms 2020". In this example, the additional pronunciation `/f O l k b a: r/` for the word "Folkbarde" caused the wrong recognition result. This pronunciation seems faulty as it canonically should contain an additional `/d/` in case of English[1] or `/d @/` in case of German[2] at the end of the phoneme sequence. The pronunciation for the correct word "verfolgbar" is contained in the baseline dictionary with the pronunciation `/f E6 f O l k b a: r/` which only differs to the "Folkbarde"-pronunciation by the additional `/f E6/` at the start.

In one interesting example, an added pronunciation lead to both and increase and a

[1] https://en.wiktionary.org/wiki/bard

[2] https://de.wiktionary.org/wiki/Barde

**Sample 5.5:** Segment 39 in "Venix - Tech News 95" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | [...] auf YouTube im Live-Stream sozusagen verfolgbar war [...] |
| baseline | [...] auf YouTube im Live-Stream sozusagen verfolgbar war [...] |
| comp-crawl | [...] auf YouTube im Live-Stream sozusagen Folkbarde war [...] |

**Sample 5.6:** Segment 24 in "Die Heldenreise des Pre-Sales Consultant" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | ich gebe dir das Mindset und die Tools um die Änderungen zu machen |
| baseline | ich gebe dir das Mainzer Tor zum die Änderungen zu machen |
| comp-crawl | ich gebe dir das Mindset und die Tools und Veränderungen zu machen |

decrease of recognition accuracy. Sample 5.6 shows the results of segment 24 in "Die Heldenreise des Pre-Sales Consultant" from test set "Anglicisms 2020". While the baseline dictionary only contained the pronunciation `/m I n t z E t/` for the word "Mindset", the comp-crawl dictionary included the additional pronunciation `/m aI n t z E t/` which lead to the correct word recognition. Similar to the last example, this missing word lead to a different choice in recognized words by the baseline model. But in this case, the correct recognition of the phrase "Mindset und die Tools" also lead to two subsequent word errors. Since the comp-crawl dictionary did not contain any additional pronunciations for the relevant words, those word errors were most likely caused by different WFST weights for the language model caused by the changed meaning of the sentence.

## 5.5 Anglicism Recognition Results

Table 5.10 on the following page shows the number of recognized anglicism entities and the calculated EER of all models created with the comparative approach by applying the test set "Anglicisms 2020". Model comp-0.5 recognized the most anglicisms (Δ 22), followed by models comp-0.75 and comp-1 (Δ 15). Interestingly, the training data for the respective P2P model used in comp-0.5 contained an exact mix of G2P and AM pronunciations. This implies that both the G2P and the acoustic model provided helpful cues to achieve the best performance in anglicism recognition. The EER results for the specific audio files are shown in Table A11 on page 123.

While all comparative models created with the anglicism list performed better in terms of anglicism recognition than the baseline, comp-crawl recognized 3 anglicisms less than the baseline did. This observation also complies to the WER results which were better than the baseline for test sets "German Broadcast 2020" and "Challenging Broadcast 2018", but worse for "Anglicisms 2020" which specifically contains segments containing anglicisms.

| Model | Recognized Entities | EER (%) |
|---|---|---|
| baseline | 824 | 39.50 |
| comp-0 | 829 | 39.13 |
| comp-0.25 | 837 | 38.55 |
| comp-0.5 | **846** | **37.89** |
| comp-0.75 | 839 | 38.40 |
| comp-1 | 839 | 38.40 |
| comp-crawl | 821 | 39.72 |

**Table 5.10:** EERs for the baseline and comparative models after applying the test set “Anglicisms 2020”.

Chapter 6

# Approach 3: Using Multitask Learning for Anglicism Detection

**Contents**


6.1 Data Collection . . . . . 73
6.2 Implementation . . . . . 74
6.3 Tuning . . . . . 76
6.4 Model Selection . . . . . 79
6.5 Evaluation . . . . . 82
6.6 Anglicism Recognition Results . . . . . 85


As anglicisms in the German language are of English heritage, they have different linguistic features from a typical German word. For example, the grapheme combinations differ from those used in German which leads to different pronunciation rules. A pronunciation by German rules based on the grapheme sequence “downloaden” could be something like [ˌdoːnloˈaːdn̩], but the correct pronunciation according to Duden[1] is [ˈdaʊ̯nlo̯ʊ̯dn̩]. This means that unusual combinations in a grapheme sequence can be used to determine if a word is an anglicism or not, hence resulting in different pronunciation rules. Based on a Seq2Seq G2P model, the anglicism classification can be added as a second task, making it an MTL model. It is expected that this will help the model understand that anglicisms are pronounced differently than “normal” German words, resulting in different phoneme conversions in case a word is classified as an anglicism.

As a preparation for this approach, a Seq2Seq G2P model will be created, trained, and evaluated against the currently used Sequitur G2P model since the latter does not support MTL. To achieve multitasking, an additional binary classification task that determines whether a word is an anglicism or not will be added to the model.

## 6.1 Data Collection

PHONOLEX core was used as training and validation data to train Fraunhofer IAIS’ German Sequitur G2P model. No test set was derived from the PHONOLEX core data since the models performance is tested in the ASR systems environment. The exact same train and

[1] https://www.duden.de/rechtschreibung/downloaden#aussprache

validation sets from the Sequitur G2P training could be retrieved, making it possible to exactly evaluate the Seq2Seq models G2P task performance against the baseline Sequitur G2P model. The train set contained 62,427 entries, the validation set 3,000. Based on the split anglicism list, all anglicisms within the two data sets were marked with a 1 while non-anglicisms were marked with a 0. The train set contained 1,388 anglicisms (2.22 %) and the validation set contained 59 anglicisms (1.96 %).

In the first tests of the MTL Seq2Seq model, the classifier did not detect any anglicisms. This was most probably caused by the small number of data with a positive class versus a very large number of data with a negative class, e.g. 2.22 % positive versus 97.78 % negative classes for the train set. To compensate the class imbalance, the wiki-v2 pronunciation dictionary from the Wiktionary approach was proportionately added to the train and validation set. With this data added, the train set contained a total of 71,102 entries with 10,063 anglicisms (16.11 %) and the validation set contained a total of 3,457 entries with 516 anglicisms (17.20 %).

Additionally, down sampled data sets have been created that offer a 50/50 anglicism class balance. Based on the combined PHONOLEX core and Wiktionary data, 61,039 entries with negative anglicism classes from the train set and 2,425 entries from the validation set have been deleted to match the number of data with positive anglicism classes. The finished down sampled train set contained a total of 20,126 entries and the down sampled validation set contained a total of 1,032 entries, both with a positive anglicism class rate of 50 %.

## 6.2 Implementation

The Seq2Seq G2P model was build after the “encoder-decoder LSTM” by Yao and Zweig (2015) which was described in Section 2.1.2 on page 18. Since Phan (2017) already published a notebook implementing Yao and Zweigs model using PyTorch, their code was used as a basis for this approach. In the code, an adaptive learning rate is used that decays when no improvements in the validation loss are observed within the last five checks. An early stopping mechanism sets in when the learning rate drops below 0.00001. Some functional and structural modifications were applied to the code while testing it on the CMUdict data. The classification task has been added as an additional task after the encoder step, transforming the single task encoder-decoder LSTM model to an MTL model (see Figure 6.1 on the next page).

The classifier is based on a binary classification example by Pascual (2018). It consists of two hidden layers and an output layer. The 500 dimensional cell state $c$ and cell output $h$ which result from the encoder are combined and used as an input for the classification task. The first hidden layer is a 1,000 dimensional linear layer with a 100 dimensional

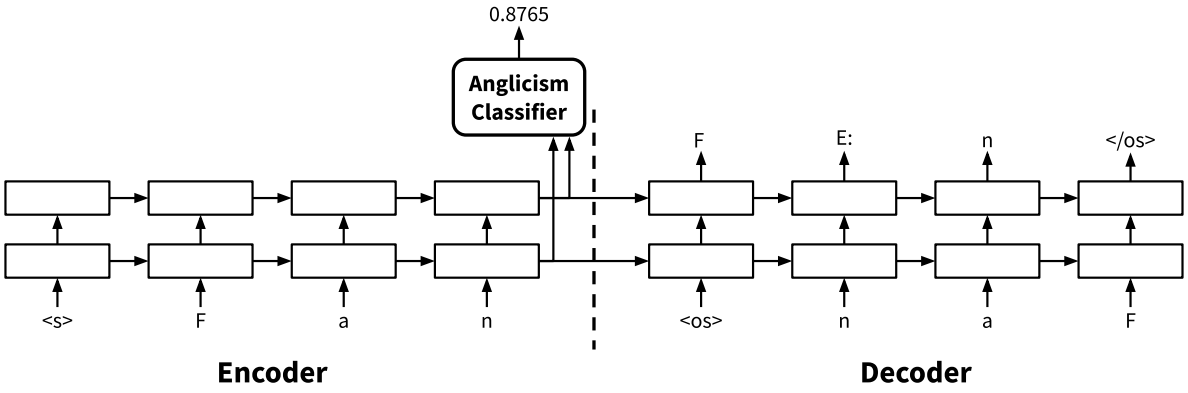


**Figure 6.1:** MTL G2P model representation for the grapheme sequence ⟨Fan⟩. The grapheme sequence is processed by the encoder that passes the output to both the decoder and the anglicism classification task. Based on the encoder output, the decoder generates the pronunciation while the classification task asserts the probability for the grapheme sequence being an anglicism. (Yao and Zweig, 2015, adapted from)

output. The ReLU function[1] is used as activation function. A dropout of 0.2 is applied to prevent overfitting. The second hidden layer is a 100 dimensional linear layer with equal output using the PReLU function[2] (He et al., 2015) with a constant $\alpha = 1$ as activation function. The output layer is a a 100 dimensional linear layer with 1 output neuron. The Sigmoid function[3] is applied to get an output value between 0 and 1. The closer the output value is to 1, the more likely it is that the word is an anglicism.

For the G2P decoder, LogSoftmax[4] was used as output activation function in the output layer. The loss was calculated with the negative log likelihood[5] since it usually goes in combination with the softmax function (Miranda, 2017). The classifier loss was calculated with the binary cross entropy[6] as this fits best with a binary classifier with an output value between 0 and 1 (Brownlee, 2019). Both losses have been combined to one total loss value in the training and validation phase to optimize on both tasks:

$$\text{Total Loss} = \text{G2P Loss} + \text{Classification Loss} \tag{6.1}$$

[1] https://pytorch.org/docs/stable/generated/torch.nn.ReLU.html

[2] https://pytorch.org/docs/stable/generated/torch.nn.PReLU.html

[3] https://pytorch.org/docs/stable/generated/torch.nn.Sigmoid.html

[4] https://pytorch.org/docs/stable/nn.functional.html#log-softmax

[5] https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html

[6] https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html

| Name | Batch Size | Epochs | Iterations/ Epoch | PER (%) | WER (%) |
|---|---|---|---|---|---|
| P1 | 1 | 5 | 62427 | 8.67 | 39.87 |
| P2 | 100 | 25 | 625 | **7.65** | **30.07** |

**Table 6.1:** PERs and WERs for the batch size configurations mentioned in Yao and Zweig (2015).

## 6.3 Tuning

The German Sequitur G2P model at Fraunhofer which is used as a baseline already performs well for German vocabulary, making it hard for a Seq2Seq model that has been implemented in a limited time frame to compete. Tested on the PHONOLEX core validation set, the baseline model showed a PER of 2.59 % and a WER[1] of 13.96 %. To somehow get close to those rates, the Seq2Seq model had to be improved. For a true comparison, the original PHONOLEX core train and validation data used to train the Sequitur G2P have been applied for the tuning.

The Seq2Seq G2P model was based on the implementation of Yao and Zweig (2015) and hence was not intended to be modified. However, they used different batch sizes for different data sets in their encoder-decoder LSTM which implies that the batch size is dependent on the training data. Therefore, tuning for this parameter has been performed to find the best configuration for the training data.

### 6.3.1 Paper Configurations

First, the two different batch sizes stated in Yao and Zweig (2015) were chosen and evaluated. In their study, they chose a batch size of 1 for the CMUdict data and a batch size of 100 for the NetTalk and Pronlex data sets. Both batch sizes were were picked as starting values for a first rough estimation of the optimal batch size for the PHONOLEX core data. Table 6.1 shows the PER and WER results. Using a batch size of 100 provided better results both in PER ($\Delta$ 1.02 %) and WER ($\Delta$ 9.80 %). However, the error rates still did not come close to the baseline results, showing a difference of 5.06 percentage points in PER and a difference of 16.11 percentage points in WER.

### 6.3.2 Manual Configurations

Additionally, manual configurations have been tested using the binary search strategy. The start value was 50 as it is the middle value between 1 and 100. The respective middle values were chosen dependent on the tuning results, resulting in a total of seven

[1]WER in context of a G2P model, defined as the ratio of hypotheses pronunciations that contain at least one wrong phoneme.

| Name | Batch Size | Epochs | Iterations/ Epoch | PER (%) | WER (%) |
|---|---|---|---|---|---|
| M10 | 10 | 6 | 6243 | 5.50 | 21.80 |
| M20 | 20 | 10 | 3122 | 5.75 | 21.60 |
| M25 | 25 | 11 | 2498 | **5.13** | 21.87 |
| M30 | 30 | 11 | 2081 | 5.46 | 22.60 |
| M40 | 40 | 15 | 1561 | 5.65 | 22.20 |
| M50 | 50 | 17 | 1249 | 5.25 | **20.80** |
| M75 | 75 | 24 | 833 | 7.86 | 27.13 |

**Table 6.2:** PERs and WERs for the manual batch size configurations.

configurations. Table 6.2 shows the PER and WER results for all manual batch size configurations. Using a batch size of 50 resulted in a PER of 5.25 % and a WER of 20.80 %. Compared to the best results in Table 6.1 on the preceding page (P2), the PER was improved by 2.40 percentage points and the WER was improved by 9.27 percentage. Using a batch size of 75 showed that the error rates increased again, being even higher than the P2 results. Using a batch size of 25 lead to another improvement, showing a PER of 5.13 % which is the top result for all configurations so far. Testing the batch size values between 25 and 50 (M30 & M40) and lower than 25 (M10 & M20) lead to no further improvements. Overall, the PER and WER values for the batch sizes between 10 and 50 are very close to each other. Looking at M25 which showed the best PER, the performance got closer to the baseline model, narrowing the difference in PER to 2.54 percentage points and the difference in WER to 7.91 percentage points.

### 6.3.3 Hyperparameter Optimization with Optuna

To also apply a more automated tuning approach, the hyperparameter optimization framework *Optuna*[1] was used to find the best configuration for the model. In Optuna, a study is created that includes several trials based on given parameter choices. Each trial performs a model training using a different parameter combination. For this tuning approach, the learning rate has been added as an additional tuning parameter. Goyal et al. (2017) established the Linear Scaling Rule[2] which states that the learning rate should be increased accordingly when increasing the minibatch size. This implies that when using a bigger batch size, a higher learning rate is needed. Adding the learning

[1] `https://optuna.org/`

[2] "**Linear Scaling Rule:** When the minibatch size is multiplied by k, multiply the learning rate by k." (Goyal et al., 2017)

rate as a second tuning parameter next to the batch size will let the Optuna framework adapt both parameter values to each other for finding the best fitting combination.

Based on the results from the manual models (see Table 6.2 on the previous page), a range of 1–50 has been chosen for the batch size. For the learning rate, a range of 0.01–0.001 has been configured. The TPE sampler (Optuna, 2018) was used to choose the parameters for the trials. This sampler uses the Tree-structured Parzen Estimator (Bergstra et al., 2011) algorithm:

> On each trial, for each parameter, TPE fits one Gaussian Mixture Model (GMM) $l(x)$ to the set of parameter values associated with the best objective values, and another GMM $g(x)$ to the remaining parameter values. It chooses the parameter value $x$ that maximizes the ratio $l(x)/g(x)$.
>
> **Optuna (2018)**

Additionally, a pruner was used to perform early stopping on unpromising trials. The hyperband pruner[1] was chosen which runs multiple instances of Successive Halving pruners in so called brackets, each processing a part of the trials. A Successive Halving pruner uses the Asynchronous Successive Halving algorithm that asynchronously trains and evaluates $n$ configurations in parallel and only pursues the top $\frac{1}{n}$ configurations (Li et al., 2020a). This process is repeated until a finite budget $B$ is reached, limiting the resources to $\frac{B}{n}$ which leads to a trade-off between $B$ and $\frac{B}{n}$ (Li et al., 2018). This trade-off is tackled by the Hyperband pruner:

> Hyperband [...] addresses this "$n$ versus $B/n$" problem by considering several possible values of $n$ for a fixed $B$, in essence performing a grid search over feasible value of $n$. Associated with each value of $n$ is a minimum resource $r$ that is allocated to all configurations before some are discarded; a larger value of $n$ corresponds to a smaller $r$ and hence more aggressive early-stopping.
>
> **Li et al. (2018)**

[1] https://optuna.readthedocs.io/en/stable/reference/generated/optuna.pruners.HyperbandPruner.html

| Name | Batch Size | Learning Rate | Epochs | Iterations/ Epoch | PER (%) | WER (%) |
|---|---|---|---|---|---|---|
| O1 | 35 | 0.009599 | 13 | 1784 | 5.95 | 20.77 |
| O2 | 30 | 0.009225 | 14 | 2081 | 5.64 | 21.90 |
| O3 | 35 | 0.009991 | 11 | 1784 | 7.48 | **19.93** |
| O4 | 22 | 0.007835 | 9 | 2838 | 5.13 | 20.53 |
| O5 | 42 | 0.005852 | 17 | 1487 | **4.96** | 21.17 |

**Table 6.3:** Batch size and learning rate configurations with their resulting PER and WER values of the top five Optuna trial models.

The Optuna study consisted of 20 trials that were optimized by their validation loss. Table 6.3 shows the parameter configurations and the respective PER and WER of the top five Optuna combinations. O3 and O5 both show better PER and WER results than the best manual combinations. O5 resulted in the best PER (4.96 %), narrowing the difference to the baseline model to 2.37 percentage points.

## 6.4 Model Selection

The two configurations with the best PER results from the manual (M25) and the Optuna tuning (O5) have been chosen to be used for the anglicism dictionary creation. Table 6.4 on the next page shows the resulting Seq2Seq MTL models. For those combinations, the three different data sets explained in Section 6.1 on page 73 have been applied: PHONOLEX core (M25-P & O5-P), PHONOLEX core combined with wiki-v2 (M25-W & O5-W) and the downsampled PHONOLEX core combined with wiki-v2 with equal class balance (M25-DS & O5-DS). Additionally, one variation was tested where the loss of the G2P task was weighted more heavily than the loss of the classification task, setting the training focus more on the G2P task (M25-WL & O5-WL). Using an $\alpha$ parameter of 0.7, the total loss calculation was changed as follows:

$$\text{Total Loss} = \alpha \cdot \text{G2P Loss} + (1 - \alpha) \cdot \text{Classification Loss} \tag{6.2}$$

Table 6.5 on the next page shows the resulting PER, WER as the G2P task evaluation metrics and the accuracy, precision, recall and F1 score as the classification task evaluation metrics after applying the respective validation set. The accuracy shows the rate of total correct classifications, the precision shows the rate of true anglicisms among all predicted anglicisms and the recall shows the rate of true anglicisms that have actually been predicted as one. The F1 score is a metric that harmonizes the precision and recall values and hence only focuses on the true anglicism classifications. Compared to the accuracy, the F1 score

| Name | Data Source | Epochs | Iterations/ Epoch |
|---|---|---|---|
| M25-P | PHONOLEX core | 7 | 2498 |
| M25-W | PHONOLEX core & wiki-v2 | 6 | 2845 |
| M25-WL | PHONOLEX core & wiki-v2 (weighed loss) | 7 | 2845 |
| M25-DS | Downsampled PHONOLEX core & wiki-v2 | 16 | 806 |
| O5-P | PHONOLEX core | 10 | 1487 |
| O5-W | PHONOLEX core & wiki-v2 | 8 | 1693 |
| O5-WL | PHONOLEX core & wiki-v2 (weighed loss) | 9 | 1693 |
| O5-DS | Downsampled PHONOLEX core & wiki-v2 | 22 | 480 |

**Table 6.4:** Selected Seq2Seq MTL models with their data source, number of epochs and number of iterations per epoch. The data source was used to create the train and validation sets.

| | G2P Task | | Anglicism Classification Task | | | |
|---|---|---|---|---|---|---|
| Name | PER | WER | Accuracy | Precision | Recall | F1 Score |
| M25-P | 5.68 % | 24.43 % | 98.03 % | 0.00 % | 0.00 % | 0.00 % |
| M25-W | 8.63 % | 30.89 % | 91.24 % | 80.69 % | 54.26 % | 64.89 % |
| M25-WL | **7.87 %** | **28.03 %** | **92.42 %** | 86.71 % | 58.14 % | 69.61 % |
| M25-DS | 11.21 % | 39.63 % | 88.66 % | **90.30 %** | **86.63 %** | **88.43 %** |
| O5-P | 6.56 % | 25.17 % | 98.03 % | 0.00 % | 0.00 % | 0.00 % |
| O5-W | 9.26 % | 31.53 % | **91.35 %** | 90.94 % | 46.71 % | 61.72 % |
| O5-WL | **7.80 %** | **27.77 %** | 91.18 % | 89.22 % | 46.51 % | 61.15 % |
| O5-DS | 11.13 % | 38.47 % | 85.76 % | **92.03 %** | **78.29 %** | **84.61 %** |

**Table 6.5:** Selected MTL models with their G2P task and anglicism classification task evaluation metrics. For models M25-P and O5-P, the precision, recall and F1 score values are 0.00 % because they did not yield any positive classifications.

is a more reliable metric when having unbalanced classes because the accuracy looks at the total correct classifications and hence can result in a good value when only choosing the majority class. (Brownlee, 2020)

The result of the classification task was a value between 0 and 1 that represents the probability for the word being an anglicism. The results have been rounded to the nearest integer to be able to calculate and compare the classification task evaluation metrics. If the value was $> 0.5$, it was rounded down to 0; if the value was $\geq 0.5$, it was rounded up to 1. In the future, this cut-off value could also be optimized.

Figure A1 on page 111 and Figure A2 on page 112 show the confusion matrices for all MTL models which contain the anglicism classification results. As mentioned in Section 6.1 on page 73, using the PHONOLEX core data did not yield any positive classification results for the anglicism detection, proving that this data alone is unfit for the anglicism MTL approach. Therefore, the PHONOLEX core models M25P and O5P will be disregarded. However, since the same PHONOLEX core train and validation sets were applied as in the isolated Seq2Se2 G2P models (see Table 6.2 on page 77 and Table 6.3 on page 79), it was observed that the PER and WER values got slightly worse, implying that the additional classification task impacted the performance of the G2P task. Comparing the training statistics, it shows that the versions without classification task were trained for 4 (M25) and 7 (O5) more epochs respectively. The G2P tasks in the MTL models have hence not reached their optima yet. Since the validation loss was influenced by the classification loss as well, it caused the early stopping mechanism to set in earlier than it did before, causing the lower PER and WER results. In the future, other early stopping criteria could be tested to further optimize the training process.

Both model M25-WL and O5-WL that implemented the additional weighted loss for the G2P task showed the best PER and WER results among the respective results. Comparing those results to the ones with the same data, but without the weighted loss (M25-W & O5-W) shows that the heavier focus on the G2P loss lead to better results for the G2P metrics. F1 score of M25-WL improved as well compared to model M25-W even though the classification loss did not contribute as much into the total loss. The classification metrics of O5-WL compared to O5-W only decreased slightly, having differences of lower than 0.6 percentage points. Looking at the confusion matrices in Figure A1 on page 111, the number of false positives and false negatives for M25-WL was decreased by 0.61 and 0.58 percentage points respectively. For the O5-WL model, the false positives decreased slightly by 0.03 percentage points, but the false negatives increased by 0.15 percentage points, causing the slightly worse values in the classification metrics.

The models M25-DS and O5-DS that were trained with the downsampled data showed the best precision, recall and F1 score values. Even though both did not have the best accuracy compared to the other models with the same configuration, the higher F1 score in particular implies that the detection of positive anglicism classifications worked better. The confusion matrices in Figure A1 on page 111 show the relatively balanced class distribution by having almost as many true positives as there are true negatives. Both M25-DS and O5-DS actually show higher relative values in their false positives and false negatives than their WL model counterparts, explaining the decrease in accuracy.

Two Seq2Seq MTL models were chosen from each configuration (M25 & O5) to create an anglicism dictionary that will be used in a dedicated ASR model. Both the M25-WL

and the O5-WL models were chosen because they showed the best G2P task results. Additionally, the M25-DS and O5-DS models were chosen since they were trained on data with balanced class distribution, so the models might have had a better basis for learning the differences between anglicisms and native German words. Based on the split anglicism list, a pronunciation dictionary was created by each of the four models.

## 6.5 Evaluation

Four dedicated ASR models were created that included the anglicism dictionary created of the respective MTL model. The dictionaries for models mtl-m25-ds, mtl-m25-wl and mtl-o5-wl contained 18,917 entries. The dictionary for model mtl-o5-ds is missing the entry for the word "a" because it was not able to generate a corresponding phoneme sequence. Therefore, the dictionary for mtl-o5-ds contains 18,916 entries.

Table A12 on page 124 shows 20 example entries from the respective anglicism dictionaries. Overall, the MTL dictionary pronunciations fit quite well judging from the examples with model mtl-o5-ds generating the most realistic results. For nine words, all models contain the exact same pronunciation. For four words, the different pronunciations sound like pronunciation variations, e.g. `/tS a r t @ r @/` and `/S a r t @ r @/` for the word "chartere" or `/r O U d S O U s/` and `r o: tS o: s/` for the word "Roadshows". For seven words, there was at least one pronunciation that did not sound like a realistic anglicism pronunciation. For the word "VIPs", only model mtl-o5-ds contained the fitting pronunciation `/v I p s/` which even corresponds to the Wiktionary pronunciation. In contrast, models mtl-m25-ds, mtl-m25-wl and mtl-o5-wl contain the pronunciations `/v i: p s/`, `/f aU p s/` and `/f aU Q i: p s/` respectively which are no typical pronunciations for this word. Another interesting example is the word "Crowdfundings" for which all models contain a different pronunciation. While model mtl-o5-ds contains the most realistic pronunciation `/k r aU d f a n d I N s/`, all other models include a mapping from ⟨u⟩ to `/U/` instead of the more fitting `/a/` with model mtl-m25-wl even containing another unfitting mapping from ⟨ow⟩ to `/o:/` instead of `/aU/`. The resulting pronunciations are `/k r aU d f U n d I N s/` (mtl-m25-ds), `/k r o: t f U n d I N s/` (mtl-m25-wl) and `/k r aU t f U n d I N s/` (mtl-o5-wl).

Table 6.6 on the following page shows the WERs of the five comparative ASR models compared to the baseline model. For the test set "Anglicisms 2020", all models performed better than the baseline with mtl-m25-wl showing the best results (Δ 0.15 %). For test sets "German Broadcast 2020" and "Challenging Broadcast 2018", however, all models showed increased WERs compared to the baseline results. Though the differences are small, ranging from 0.01–0.04 percentage points for "German Broadcast 2020" and 0.02–0.06 percentage points for "Challenging Broadcast 2018", it shows that some of the

| Model | Anglicisms 2020 (%) | German Broadcast 2020 (%) | Challenging Broadcast 2018 (%) |
|---|---|---|---|
| baseline | 15.80 | **6.56** | **10.84** |
| mtl-m25-ds | 15.67 | 6.60 | 10.90 |
| mtl-o5-ds | 15.73 | 6.60 | 10.90 |
| mtl-m25-wl | **15.65** | 6.57 | 10.86 |
| mtl-o5-wl | 15.73 | 6.59 | 10.86 |

**Table 6.6:** WERs for the baseline and the MTL ASR models.

**Sample 6.1:** Segment 1 in "Wirtschaft regional - Teeherstellung in Bremen" from test set "Challenging Broadcast 2018".

| | |
|---|---|
| **reference** | `[...] der Tee boomt ja seit einigen Jahren [...]` |
| baseline | `[...] der Tee Boom der seit einigen Jahren [...]` |
| mtl-m25-ds | `[...] der the Boom der seit einigen Jahren [...]` |
| mtl-o5-ds | `[...] der the Boom der seit einigen Jahren [...]` |
| mtl-m25-wl | `[...] der Tee Boom der seit einigen Jahren [...]` |
| mtl-o5-wl | `[...] der Tee Boom der seit einigen Jahren [...]` |

generated anglicism pronunciations have negatively influenced the recognition results for those two typical German test sets. The best model from the MTL approach is mtl-m25-wl. Table A13 on page 125, Table A14 on page 126 and Table A15 on page 127 show the detailed comparison of the results for the specific audio files.

Sample 6.1 shows an extract of segment 1 in "Wirtschaft regional - Teeherstellung in Bremen" from test set "Challenging Broadcast 2018". Here, mtl-m25-ds and mtl-o5-ds both recognized the English word "the" instead of the German word "Tee". Both anglicism dictionaries list the word "the" with the pronunciation `/t e:/` which actually differs from the canonical pronunciation[1]. In the baseline dictionary, the word "Tee" is listed with the same pronunciation, making them homophones. Supposedly, the language model saw a better fit in the combination "the Boom" than it did for "Tee Boom", hence choosing "the" over "Tee" for the models that listed the respective pronunciation in their dictionaries. However, this example shows that quality issues in the G2P model might cause recognition issues in the ASR results.

Sample 6.2 on the next page shows an extract of the recognition results of segment 13 in "Besuch von Bundeskanzlerin Merkel in Indien" from test set "German Broadcast

[1] `https://en.wiktionary.org/wiki/the`

**Sample 6.2:** Segment 13 in "Besuch von Bundeskanzlerin Merkel in Indien" from test set "German Broadcast 2020".

| | |
|---|---|
| **reference** | denn der indische Markt biete enorme Chancen [...] |
| baseline | denn der indische Markt biete enorme Chancen [...] |
| mtl-m25-ds | denn der indische Markt Beate enorme Chancen [...] |
| mtl-o5-ds | denn der indische Markt Beate enorme Chancen [...] |
| mtl-m25-wl | denn der indische Markt Beate enorme Chancen [...] |
| mtl-o5-wl | denn der indische Markt Beate enorme Chancen [...] |

**Sample 6.3:** Segment 57 in "Rezo - Die Zerstörung der Presse" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | [...] völligen Bullshit überzeugt als Wahrheit rüberzubringen |
| baseline | [...] völligen Wohlstand überzeugt als Wahrheit rüberzubringen |
| mtl-m25-ds | [...] völligen Bullshit überzeugt als Wahrheit rüberzubringen |
| mtl-o5-ds | [...] völligen Bullshit überzeugt als Wahrheit rüberzubringen |
| mtl-m25-wl | [...] völligen Bullshit überzeugt als Wahrheit rüberzubringen |
| mtl-o5-wl | [...] völligen Bullshit überzeugt als Wahrheit rüberzubringen |

2020" which is another example for a negative influence caused by the supplementary anglicism dictionary. The German word "biete" has been falsely recognized as "Beate" by all MTL models. It was surprising seeing this word in the recognition results because it is mainly known as a German female given name and also not considered an inflection of the anglicism "Beat"[1]. However, when scrolling down on the respective Wiktionary page, there is another meaning listed for the word "Beat" as a German male given name. This entry also has an own inflection table. Since no distinction for parts of speech were made when crawling the German inflection tables, the inflections for this proper noun were crawled as well when adding the Wiktionary inflections to the anglicism list (see Section 3.2.2 on page 33).

For "Beate", all MTL dictionaries contain the pronunciation `/b i: t @/` which is the same as listed for "biete" in the baseline dictionary. Looking at the Wiktionary pronunciation for "beate"[2], the MTL models actually generated the correct pronunciation when assuming this was an anglicism. Also, "Beate" was contained in the training data with a positive anglicism class. This example shows that even though the G2P performed well when generating this pronunciation, it was an error in the data that caused this word error, showing that data quality is crucial when creating a G2P model.

[1] https://de.wiktionary.org/wiki/Beat

[2] https://de.wiktionary.org/wiki/beate

**Sample 6.4:** Segment 51 in "Rezo - Die Zerstörung der CDU" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | ist doch nice |
| baseline | ist doch naiv |
| mtl-m25-ds | ist doch nice |
| mtl-o5-ds | ist doch naiv |
| mtl-m25-wl | ist doch naiv |
| mtl-o5-wl | ist doch naiv |

While all models performed slightly worse for test sets "German Broadcast 2020" and "Challenging Broadcast 2018", it did show improved WERs for "Anglicisms 2020" compared to the baseline model. Sample 6.3 on the previous page shows segment 57 in "Rezo - Die Zerstörung der Presse" from test set "Anglicisms 2020" as a positive example for additional anglicism pronunciations in the dictionary. The word "Bullshit" was not recognized by the baseline model, but by all MTL models. This is caused by a faulty pronunciaiton in the baseline dicitonary. Here, "Bullshit" is listed with the pronunciation /b U l s h I t/ which mapped the grapheme ⟨sh⟩ to /s h/ instead of /S/. The correct pronunciation /b U l S I t/ is contained in the supplementary anglicism dictionaries for all four MTL models which caused the word "Bullshit" to be recognized correctly.

Another positive example is shown in Sample 6.4 which shows segment 51 in "Rezo - Die Zerstörung der CDU" from test set "Anglicisms 2020". The word "nice" was only recognized by mtl-m25-ds while all other models recognized "naiv" instead. Looking at "nice" in the pronunciation dictionaries, the baseline, mtl-o5-ds, mtl-m25-wl and mtl-o5-wl dictionaries only list the pronunciaiton /n I s/. Only the mtl-m25-ds dictionary lists the pronunciation /n aI s/ which corresponds to the canonical pronunciation[1]. For the word "naiv", the baseline dictionary lists the pronunciation /n aI f/ which caused the word error. This was an interesting observation because the huge baseline dictionary is missing the canonical pronunciation for "naiv". Instead of pronouncing the diphthong /aI/, the two vowels are pronounced separately, leading to the pronunciation /n a i: f/ [2]. It should be investigated if the baseline model is actually able to recognize the word "naiv".

## 6.6 Anglicism Recognition Results

Table 4.5 on page 50 shows the number of recognized anglicism entities and the calculated EER of all models created with the MTL approach by applying the test set "Anglicisms

[1] https://www.duden.de/rechtschreibung/nice#aussprache

[2] https://de.wiktionary.org/wiki/naiv

| Model | Recognized Entities | EER (%) |
|---|---|---|
| baseline | 824 | 39.50 |
| mtl-m25-ds | 839 | 38.40 |
| mtl-o5-ds | 838 | 38.47 |
| mtl-m25-wl | **840** | **38.33** |
| mtl-o5-wl | 834 | 38.77 |

**Table 6.7:** EERs for the baseline and MTL models after applying the test set "Anglicisms 2020".

2020". Model mtl-m25-wl recognized the most anglicisms (Δ 16) and hence shows the best EER value. The EER results for the specific audio files are shown in Table A16 on page 128.

It was expected that the models trained with downsampled data (ds models) would perform better than the models trained with weighted losses (wl models) because they were trained on more balanced training data. While both ds models only performed slightly worse than mtl-m25-wl having missed 1 (mtl-m25-ds) and 2 (mtl-o5-ds) more anglicisms, mtl-o5-wl performed worse by having missed more 6 anglicisms. Looking at the evaluation metrics of the corresponding G2P models (see Table 6.5 on page 80), O5-WL showed the worst F1 score of all selected models (61.15 %), so it might not have classified the anglicisms as well as the other models, hence generating unfitting pronunciations for them. Model mtl-m25-wl showed the best anglicism recognition performance even though the corresponding G2P model was trained with a worse class balance in the training data than both the ds models. However, G2P model generating the anglicism pronunciations for mtl-m25-wl showed lower PER and WER values than the G2P models used for creating the anglicism pronunciations for the ds models. It is unclear of the lower performance by the ds models was caused by the performance of the classification task or the G2P task as the latter could have caused correctly detected anglicisms to get an unfit pronunciation. If the generated pronunciation does not fit the canonical pronunciation, the word cannot be recognized regardless of the correct anglicism classification it might have made.

# Chapter 7

# Discussion

After all experiments have been concluded, the results were compared against each other to find out which model performed best. Since test set "Anglicisms 2020" was specifically created to evaluate anglicism recognition, it is used to determine each models anglicism recognition performance. Table 7.1 on the following page shows the number of entries in the anglicism pronunciation lexicon, the number of total entries in the pronunciation dictionary as well as the WER and EER values for test set "Anglicisms 2020" for all ASR models. Also, a mean WER has been calculated to evaluate the total performance among all test sets. The anglicism pronunciation dictionary refers to the dictionary that has been created with the respective approach. Combined with the baseline dictionary, the pronunciation dictionary expanded by 3.690 for the smallest anglicism dictionary (wiki-v1) and by 389,097 for the biggest anglicism dictionary (comp-crawl). Most anglicism dictionaries contained 18,917 pronunciations as this was the size of the split anglicism list.

The number of added anglicism pronunciations did not seem to correlate to the number of recognized anglicisms. Looking at the models with an anglicism dictionary size of 18,917, they all vary in their recognition values. The Spearman correlation[1] has been calculated for the anglicism dictionary sizes and the number of recognized anglicisms to objectively check if there is a correlation between the number of added anglicism pronunciations and the anglicisms that were actually recognized. The result was 0.133 which only shows a very low positive correlation between the two values, meaning that with increasing dictionary size the number of recognized anglicisms has only a low tendency to increase as well.

The EERs and WERs for test set "Anglicisms 2020", however, did seem to correlate with each other. Figure 7.1 on page 89 shows the EER and WER values of all ASR models with the WER plotted on the x-axis and the EER plotted on the y-axis. Looking at the diagram, the values seem to linearly correlate to each other. Calculating the Pearson correlation between the EERs and WERs results in a value of 0.964, meaning that there is a strong positive correlation between the two values. When the EER decreases and hence more anglicisms are recognized, the WER decreases as well because more words could correctly be recognized.

[1]The Spearman correlation has been chosen to only assert a monotonic relationship instead of a linear relationship because of the large discrepancy in the number of anglicism pronunciations versus the relatively similar numbers of recognized anglicisms.

| Model | Anglicism Dict. Entries | Total Dict. Entries | Recognized Anglicisms | Anglicisms 2020 WER (%) | Anglicisms 2020 EER (%) | Mean WER (%) |
|---|---|---|---|---|---|---|
| Baseline | 0 | 3,500,170 | 824 | 15.80 | 39.50 | 11.07 |
| wiki-base | 9,748 | 3,506,509 | 834 | 15.72 | 38.77 | 11.04 |
| wiki-v1 | 6,292 | 3,503,860 | 836 | 15.71 | 38.62 | 11.03 |
| wiki-v2 | 9,802 | 3,506,610 | 836 | 15.70 | 38.62 | 11.01 |
| comp-0 | 18,917 | 3,519,087 | 829 | 15.80 | 39.13 | 11.08 |
| comp-0.25 | 18,917 | 3,519,087 | 837 | 15.67 | 38.55 | 11.02 |
| comp-0.5 | 18,917 | 3,519,087 | **846** | **15.59** | **37.89** | **10.99** |
| comp-0.75 | 18,916 | 3,519,086 | 839 | 15.67 | 38.40 | 11.02 |
| comp-1 | 18,916 | 3,519,086 | 839 | 15.67 | 38.40 | 11.02 |
| comp-crawl | 389,119 | 3,889,267 | 821 | 15.89 | 39.72 | 11.09 |
| mtl-m25-ds | 18,917 | 3,519,087 | 839 | 15.67 | 38.40 | 11.06 |
| mtl-o5-ds | 18,916 | 3,519,086 | 838 | 15.73 | 38.47 | 11.08 |
| mtl-m25-wl | 18,917 | 3,519,087 | 840 | 15.65 | 38.33 | 11.03 |
| mtl-o5-wl | 18,917 | 3,519,086 | 834 | 15.73 | 38.77 | 11.06 |

**Table 7.1:** All ASR models with their number of anglicism pronunciation dictionary entries, number of total pronunciation dictionary entries (baseline and anglicism pronunciations), number of recognized anglicisms, their WER and EER values for test set "Anglicisms 2020" and a mean WER value that corresponds to the average WER of all test sets ("Anglicisms 2020", "German Broadcast 2020" and "Challenging Broadcast 2018").

Looking at the mean WERs, all models except for comp-0, comp-crawl and mtl-o5-ds performed better than the baseline model. Adding anglicism pronunciations to the baseline dictionary carried the risk of falsifying the recognition results due to homophones and wrongly generated phoneme sequences. However, the improved mean WERs show that in average almost all models were able to exceed the general recognition performance compared to the baseline for the three tested test sets.

The best models of each approach are wiki-v2, comp-05 and mtl-m25-wl. Table 7.2 on the following page shows five example anglicisms with their respective pronunciations from the models supplementary anglicism pronunciation dictionaries. While wiki-v2 also includes pronunciation variations if they were available in the Wiktionary entry, the models from the comparative and MTL approaches only included the respective best pronunciation in the dictionary. The example pronunciations show how different the phoneme sequence outputs from the G2P models turned out.

The wiki-v2 pronunciations in the example table can be interpreted as canonical anglicism pronunciations as they were obtained from Wiktionary directly. The pronunciations from comp-0.5 show mixed results. While some pronunciations like `/Q i: m E I l k O n t i/` and `/l aI f S t r i: m s/` seem to be valid pronunciation variations of the respective word, pronunciations like `/Q aI n t ts s E n t m Y n ts @/` and `/s m O: S s/` rather seemed like

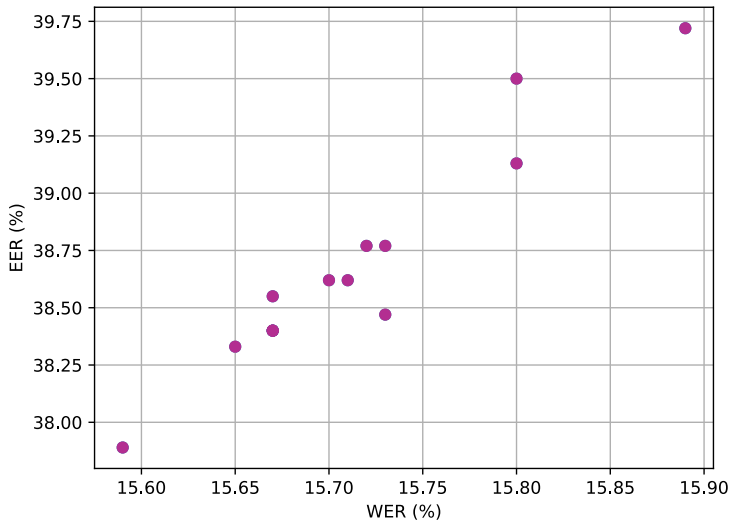


**Figure 7.1:** Correlation of WER and EER values of test set “Anglicisms 2020” for all ASR models.

| Word | wiki-v2 | comp-0.5 | mtl-m25-wl |
|---|---|---|---|
| 10-Cent-Münze | ts e: n s E n t m Y n ts @<br>ts e: n ts E n t m Y n ts @ | Q aI n t ts s E n t m Y n ts @ | ts e: ts E n t m Y n ts @ |
| E-Mail-Konti | Q i: m e I l k O n t i<br>Q i: m e: l k O n t i | Q i: m E I l k O n t i | Q i: m e: l k O n t i |
| Live-Streams | l aI f s t r i: m s | l aI f S t r i: m s | l aI f s t r i: m s |
| Smashs | s m E S s<br>s m E: S s | s m O: S s | s m E S s |
| zoome | ts u: m @<br>z u: m @ | z u: m @ | ts u: m @ |

**Table 7.2:** Five example words with their respective pronunciation(s) from the best performing pronunciation dictionary of each approach.

pronunciations for different words, in this case “1-Cent-Münze” and “Smoshs”. Pronunciations like this are dangerous as they can lead to word errors. The pronunciations from mtl-m25-wl look quite good, except for the pronunciation for the word “10-Cent-Münze”. Here, the phoneme `/n/` seems to be missing, resulting in the generated pronunciation `/ts e: ts E n t m Y n ts @/` which rather sounds like a pronunciation for the word “Zeh-Cent-Münze”. As this phenomenon has been observed multiple times for all MTL models, it seems to be an issue with the data underlying Seq2Seq G2P component.

Numbers generally cause problems in G2P models as they are pronounced differently depending on the context (e.g. “seventeen” vs. “one seven”). Therefore, the numerals are spelled out or deleted from the training data of a G2P model to avoid issues (van Hessen et al., n.d.). In the annotation of ASR model test sets, numerals are usually spelled out as

| Model | Anglicisms 2020 WER (%) | German Broadcast 2020 WER (%) | Challenging Broadcast 2018 WER (%) | Mean WER (%) |
|---|---|---|---|---|
| baseline | 15.80 | 6.56 | 10.84 | 11.07 |
| wiki-v2 | 15.70 | **6.51** | 10.83 | 11.01 |
| comp-0.5 | **15.59** | 6.57 | **10.82** | **10.99** |
| mtl-m25-wl | 15.65 | 6.57 | 10.86 | 11.03 |

**Table 7.3:** WERs of the best models from each approach, including the baseline model for comparison.

well to express the actual spoken words, so the recognition of numerals is not required by a G2P model. Unfortunately, this practice was neglected in this work, so numerals were contained in the anglicism list. In future work, numerals should either be spelled out or deleted from the data.

Table 7.3 shows the WERs for all test sets and the mean WER of the best models including the baseline for comparison. Model comp-0.5 showed the best WERs for test sets "Anglicisms 2020" (15.59 %) and "Challenging Broadcast 2018" (10.82 %) while wiki-v2 shows the best WER for test set "German Broadcast 2020" (6.51 %). Overall, comp-0.5 performed best, showing the lowest mean WER of 10.99 % with an improvement of 0.08 percentage points compared to the baseline. While comp-0.5 performed a bit worse in test set "German Broadcast 2020" than the baseline, it was able to decrease the WER values by 0.21 percentage points for "Anglicisms 2020" and by 0.02 percentage points for "Challenging Broadcast 2018". Even though the difference is small, the anglicism dictionary created with the comparative approach with an AM weight of 0.5 was able to improve the performance of the baseline model. Considering that the portion of anglicisms in German speech is estimated at 4.53 % (Hunt, 2019), this small improvement can be seen as a successful attempt on improving anglicism recognition in German ASR.

# Chapter 8

# Conclusion

**Contents**

8.1 Research Questions . . . . . 92
8.2 Challenges . . . . . 94

In this work, three approaches have been designed and tested for their performance in anglicism recognition. The experiments in those approaches resulted in 13 different ASR models. Test set "Anglicisms 2020" was specifically created for this work to evaluate the anglicism recognition. This test set only contains segments with loanwords. The respective anglicisms have been marked to measure a specific EER which evaluates the recognition of anglicisms.

While only utilizing existing pronunciations from open source data, the Wiktionary approach was able to improve the recognition of anglicisms. Luckily, the German Wiktionary is nicely maintained which makes it a good source for pronunciations in general. The best model of this approach, wiki-v2, recognized twelve more anglicisms than the baseline model. Also, it shows the second best mean WER value of all models, meaning that it generally performed well in all chosen test sets. The Wiktionary approach is hence a quick and easy way to improve anglicism recognition in German ASR.

With model comp-0.5, the comparative approach created the best ASR model in this work. The comparative approach created anglicism pronunciations based on both a German and an English G2P model. The P2P model allowed the use of English pronunciations in the German ASR system, making it possible to take advantage of the English heritage of anglicisms. The confidence measures of the resulting German and English G2P results were used to choose the respective best pronunciation. Model comp-0.5 was able to recognize 22 more anglicisms than the baseline. It also showed the best mean WER of all models, making it the best of all ASR models created in this work. While the comparative approach proved to be a good possibility for generating anglicism pronunciations on unseen data, it is highly dependent on the quality of the P2P model to create correct equivalents of English pronunciations using the German phoneme set.

The MTL approach used deep learning to solve the problem of unfit anglicism pronunciations created by a monolingual German G2P. By adding a binary classification task that determined weather a word is an anglicism or not, the model was able to treat anglicisms

differently. This way, the G2P task learned that anglicisms are pronounced differently than native German words, applying different implicit rules depending on the classification result. The best model in this approach was mtl-m25-wl which recognized 16 more anglicisms than the baseline ASR model. While it only showed the sixth best mean WER result of all models, it did show a better general performance than the baseline model. The MTL approach is a modern and efficient way to generate anglicism pronunciations since, unlike the comparative approach, only one model is needed to create both native German and anglicism pronunciations. However, better data is needed to produce results of higher quality. While an anglicism list was used to automatically classify the training data, the positive anglicism classifications were limited to the content of this list. Therefore, true anglicisms were potentially missed due to spelling differences or simply by their absence in the anglicism list.

The anglicism recognition results from the baseline ASR model was improved by twelve models with only model comp-crawl recognizing less anglicisms than the baseline did. Looking at the total performance in word errors for all three test sets, ten models were able to improve the general recognition results. Only models comp-0, comp-crawl and mtl-o5-ds showed increased mean WERs compared to the baseline. Even though the improvements were small, they can be interpreted as a success since anglicisms only make up a small part of the German language. However, the number of recognized anglicisms did not increase much. Of all 1,362 marked anglicisms, the best model comp-0.5 recognized 846 which is only 22 more than the baseline recognized.

Insufficient and inconsistent data potentially lead to a loss in quality of the resulting pronunciations (see Section 8.2 on page 94). Also, the implementation and tuning time of the G2P and P2P models was limited due to time restrictions linked to this work. The resulting anglicism pronunciations sometimes did not comply with a realistic pronunciation variation of the respective anglicism. It is expected that with better data and model implementations, the quality of the generated anglicism pronunciations can be improved which will lead to an increased number of recognized anglicisms.

## 8.1 Research Questions

Three research questions were stated at the beginning of this work. After the experiments have concluded, the research questions can be answered as follows:

#### Q1: How can anglicism recognition be improved in German ASR?

With the Wiktionary approach, the anglicism recognition performance was improved by simply adding anglicism pronunciations from external sources to the pronunciation dictionary. Thanks to those new entries, the ASR model was able to recognize more

anglicisms.

With the comparative approach, both a German and an English G2P model were used to generate anglicism pronunciations. Based on the confidence measure, the respective best resulting pronunciation for an anglicism grapheme sequence was chosen to be added to the pronunciation dictionary. To comply with the monolingual German ASR system, the winning pronunciations containing English phonemes were transformed to their German equivalents. The comparative approach produced the best performing ASR model in this work, confirming the assumption that the English heritage of anglicisms may help in generating anglicism pronunciations.

With the MTL approach, a new monolingual German G2P model was created that contained an additional classification task to distinguish anglicisms. When an anglicism was detected, the model generated the pronunciation differently than it would do for a native German word. The anglicism pronunciation could be improved using this model since it generated more realistic anglicism pronunciations.

**Q2: Can pronunciation generation of anglicisms be improved considering their English etymology?**

In the comparative approach, an additional English G2P model was involved to utilize the English heritage of anglicisms. If the German G2P model produced a pronunciation with a low confidence measure for an anglicism, the English G2P model was used instead to generate the respective pronunciation. An additional P2P model was needed to map the English phonemes to their German counterparts, preserving their English pronunciation as best as possible. The comparative approach produced the model with the best anglicism recognition and general performance compared to all other models, including the baseline. Hence, the pronunciation mapping of anglicisms was improved by considering their English etymology.

**Q3: How can anglicisms be distinguished in the German language?**

In model comp-crawl from the comparative approach, the pronunciations confidence measure was used to determine weather a word is an anglicism. Usually, a monolingual German G2P model creates the pronunciation dictionary at Fraunhofer IAIS based on webcrawl results. For model comp-crawl, all words with pronunciation results that showed a confidence measure of 0.4 or lower were processed by an English G2P model. The confidence measure of the resulting pronunciations were then compared to those of the German G2P model. If the pronunciation of the English G2P was best, it was added to the supplementary anglicism dictionary.

Model comp-crawl recognized the least anglicisms of all models including the baseline

which means that this approach variation even decreased the performance. However, the bad performance could also be caused by missing anglicisms in the web crawl results. Of the split anglicism list that was used as basis for the supplementary pronunciation dictionary in the other models, 6,312 words were not contained in the web crawl results. Hence, those words did not have a chance to get recognized compared to the other models where the anglicism pronunciations were specifically generated from this list.The performance of anglicism detection based on the confidence measure can therefore not be conclusively determined.

In the MTL approach, an additional classification task was implemented to detect anglicisms. Based on the assumption that anglicisms use different spelling rules compared to typical German words, the classifier determined if a word is an anglicism based on the grapheme sequence. Looking at the F1 scores in Table 6.5 on page 80 which shows the metrics for the underlying G2P models of ASR models mtl-m25-ds, mtl-o5-ds, mtl-m25-wl and mtl-o5-wl, the anglicism detection worked quite well for both DS models while it only yielded mediocre classification results for both WL models. The DS models were trained on downsampled data that showed a perfect class balance for the containing pronunciations. However, the downsampled data did not contain enough pronunciations to train a reliable G2P task, hence resulting in higher PER and WER values. For both WL models, the pronunciations with positive anglicism classification only accounted for 16.11 % of the total training data which could have lead to the model choosing a negative classification result more likely.

Looking at the recognized anglicism results for the respective ASR models, all models performed better than the baseline, but mtl-m25-wl turned out to recognize one more anglicisms than its ds counterpart mtl-m25-ds. Seeing that the pronunciation quality plays a role in recognizing anglicisms, the performance of the MTL approach on detecting anglicisms cannot be conclusively determined in the scope of spoken language. For the isolated written anglicisms, however, the G2P models trained with the downsampled data showed promising results that can potentially be improved by using better training data.

## 8.2 Challenges

While the conducted experiments show an overall success, some issues were also encountered in the course of this work. Solving these problems could benefit the quality of the results. The following sections describe the main challenges that have been experienced with the three different approaches.

### Wiktionary Approach

The code on the Wiktionary websites was challenging to crawl as the same elements were sometimes used in different contexts. Hence, mistakes were made in the first crawl process of which the data was used for building the first ASR model for this work. Luckily, this crawl resulted in unique pronunciations compared to the fixed crawl, so it was used regardless in form of a variation (see Section 4.3.1 on page 46). Another mistake regarding the Wiktionary data was that the dumps were not noticed until it was too late into the work. Wiktionary provides data dumps that are more easy to parse than actually crawling the website. Also, Wiktionary blocks clients for a short time when too many requests are made on their web pages, resulting in unsuccessful crawls and hence prolonging the data sourcing process.

Another issue was that some anglicism spellings had different word meanings in the German Wiktionary. An example is the word “Absence” that specifically means “absence of mind” when used in the German language. This word is actually a French loanword, hence Wiktionary lists its pronunciation as `/Q a p s a~: s/`. Another example is the word “human” for which a corresponding German word exists, resulting in actual German inflections (e.g. “humaner”, “humansten”) being included in the anglicism list. Another related issue was that Wiktionary lists “Sport” as an ancient anglicism which caused the anglicism list to contain words like “Sport-Abteilung” or “Sport-Geschichte”.

While cases like this did not cause any issues in the Wiktionary approach itself since the added pronunciations are actually correct in the context of the German language, they were problematic when it came to the resulting anglicism list. Aside from being the basis for generating anglicism pronunciations, the anglicism list was also used to mark anglicisms in the test set “Anglicisms 2020” and in the train data of the MTL approach. The entity mappings in the test set “Anglicisms 2020” have manually been checked and corrected to correctly mark all included anglicisms, but the MTL classifications were not since it was not possible to manually check over 70,000 entries in the train and test data due to the work’s time restrictions. Since the anglicism list potentially contained words with multiple meanings in the German language as shown in the examples, using this list might have led to false anglicism classifications in the MTL train and test sets.

### Comparative Approach

While the comparative approach was generally easy to implement, creating the P2P model was a challenge. Originally, it was planned to create an own TTS model based on Tacotron 2[1] to generate the audio files from the English training data. However, this attempt was dismissed due to development environment issues that took too long to resolve. Luckily,

[1] https://github.com/NVIDIA/tacotron2/

| Source | Pronunciation | Confidence |
|---|---|---|
| CMUdict | M IH D K AE P | (original) |
| German G2P | m I t k a p | 0.2689 |
| Aria | m e: t E6 p | 0.9549 |
| Salli | m I l t k E6 p | 0.9474 |
| Matthew | m I t k E6 p | 0.3712 |
| Joey | m 2: t k E6 p | 0.8003 |

**Table 8.1:** Recognition results for the word "MIDCAP" from CMUdict (Carnegie Mellon University, 2014). The CMUdict pronunciation is the original ARPABET pronunciation that the audio files of the voices were based on. Please note that the pronunciations in rows 2–6 are written in BAS-SAMPA notation.

the TTS services Amazon Polly and Microsoft Azure TTS that were able to handle IPA input could be used instead.

Four different voices were used for creating the audio training data to ensure having backups if one voice created results of lower quality. The phoneme recognition feature from the live recognizer that used the AM to generate pronunciations from the audio files was still in the beta phase, so some mistakes were expected when using the application. However, looking at the resulting data, it seemed like the confidence measure did often not correspond to the quality of the result.

Table 8.1 shows the live recognition results (BAS-SAMPA notation) for the audio files created based on the CMUdict ARPABET pronunciation `/M IH D K AE P/` for the word "MIDCAP". The example shows that the actual best result is `/m I t k E6 p/` which was created by the Matthew audio file, but it only has a confidence measure of 0.3712. If the best pronunciation was solely chosen by comparing the confidence measures, the best result would be `/m e: t E6 p/` by Aria which is not as fitting. While the unreliable confidence measures could have either been caused by a bad audio result in the source file or mistakes made by the early version of the phoneme recognition feature, it was clear that an additional criterion was needed to choose the best pronunciation. Therefore, the Levenshtein distance comparison was used for choosing the best result which is described in Section 5.1.2 on page 56. The confidence measure was only used as a second criterion in case the lowest Levenshtein distance applied to more than one word.

### MTL Approach

As already mentioned in the Wiktionary approach challenges, the anglicism list containing words with additional German meanings might have led to false anglicism declarations since it was used to automatically declare anglicisms in the PHONOLEX data. An example

for this is the word “human” which is listed with the German pronunciation `/h u m a: n/`, but got listed as an anglicism. Also, positive anglicism declarations might have been missed if a word was not contained in the anglicism list. Two examples from the MTL train data are the words “all-you-can-eat” and “yes” which have not been declared as an anglicism since they were not present in the anglicism list. This possibly made it hard for the MTL G2P models to learn distinguishing anglicisms correctly.

The false and missing anglicism classifications lead to a decrease in data quality. But even assuming the classifications were correct, the number of positive anglicism classifications in the train set was too low. 1,388 anglicisms could be found in the PHONOLEX core train data based on the split anglicism list which means that only 2.22 % of the entries were classified as anglicisms. This lead to a huge class imbalance that caused the trained MTL model to exclusively choose negative anglicism classifications. Adding the wiki-v2 pronunciations helped the classifier to actually learn a differentiation between anglicisms and native German words as the portion of positive anglicism classifications in the train data rose to 16.11 %. However, the entries in the wiki-v2 dictionary were the canonical results that the MTL model would yet have to generate since both the wiki-v2 dictionary and the MTL dictionaries are based on the same anglicism list. Hence, adding the wiki-v2 results potentially falsified the MTL results. The downsampled data intensified this problem since the wiki-v2 dictionary made up half of the train and test data respectively. While the classification results of the ds models were quite good, it is unclear if this was caused by the reviewed anglicisms already being in the train data.

# Chapter 9

# Future Work

Based on the findings obtained in this work, there are several aspects that can further be experimented in the future.

### Improving the P2P model

The P2P model that was used to map English ARPABET to German BAS-SAMPA pronunciations can further be improved. While the German Sequitur G2P model was well-trained and thoroughly evaluated by Fraunhofer IAIS, the phoneme feature of the live recognizer was still in its beta phase. With a more refined application to generate the pronunciations based on the English audio files, better data quality could be achieved which would lead to an improved P2P model. Also, other implementations like the Allosaurus phone recognizer[1] (Li et al., 2020b) could be used to create pronunciations based on the audio data. The resulting pronunciations could then be evaluated against the data created by the IAIS live recognizer to better judge the data quality.

Aside from the data, the P2P model itself could also further be improved. Like proposed by Bruguier et al. (2017), a deep learning approach could be used that both integrates a G2P task and an AM task to generate pronunciations based on the English grapheme sequences and phonetized pronunciations. Alternatively, in a more simple approach, a Seq2Seq P2P based on Yao and Zweig (2015) could be built using the same training data that was applied to train the Sequitur P2P model in Section 5.1.1 on page 53.

### Improving the MTL model

As already mentioned in Section 8.2 on page 96, the training data should be improved by fixing potentially wrong anglicism classifications in the existing data and by adding more anglicism data to offer a better anglicism class balance. Also, since only an example classifier has been used due to time constraints, the classification task should be adjusted to the MTL model by trying out different configurations using a tuning framework. The Seq2Seq G2P task itself could be further developed as well by improving the code and comparing it to models from other publications.

[1] A live demo of Allosaurus by Li et al. (2020b) can be found at `https://www.dictate.app/phone`. After pasting an audio clip, the tool will generate its respective pronunciation. When selecting “German” on the left sidebar, only phones from the German language will be used.

### Improving the Data

It was noticed that the PHONOLEX core data contains some faulty entries. For example, the word "gemma" is listed with the pronunciation `/g e: m a/`. Usually, an ⟨e⟩ followed by a double consonant is pronounced shortly, so using `/e:/` in the respective pronunciation is not correct. Instead, `/g @ m a/` would be a more realistic pronunciation. Another example is the word "Mobiltelefonss" which contains a spelling mistake with having an additional "s" at the end. Even disregarding the spelling mistake, the provided pronunciation `/m o b I l t e: l @ f O n s/` does not seem to fit. For the correct spelling "Mobiltelefons", it lists the realistic pronunciation `/m o b i: l t e l e f o: n s/` which differs in 4 phonemes, all being vocals, from the "Mobiltelefonss" pronunciation. The additional "s" at the end could not have caused such a grave difference in pronunciations. A procedure could be designed that checks the PHONOLEX core data for such mistakes. The grapheme sequences could be checked against a German dictionary like Duden or Wiktionary to filter out possible misspellings. To filter out potentially faulty pronunciations, a well-trained G2P could be used to produce pronunciations based on the PHONOLEX core grapheme sequences that will then be compared to the PHONOLEX phoneme sequences. All words and pronunciations that have been filtered out using those methods should be manually checked before they are discarded from the train and test data. This procedure is only a simple example, but by cleaning the PHONOLEX core data from faulty entries, the quality of the resulting G2P model could be improved.

For creating the anglicism list that was the basis for building a supplementary anglicism pronunciation dictionary, the Wiktionary anglicism indices as well as the VDS Anglizismenindex have been used. This way, 18,967 anglicisms could be derived. By further expanding the anglicism list, more anglicism pronunciations could be generated, increasing the chance of recognizing more anglicisms in German ASR. To achieve this, more resources like Görlach (2005) could be researched and used to expand the existing anglicism list. Also, anglicisms could be derived from the Fraunhofer IAIS web crawls using a method established by Coats (2019) to frequently retrieve up-to-date anglicism data. Coats build a non-standard German verbal anglicism corpus based on data from the social media platform Twitter. First, he selected 36,240,530 German Tweets that were tokenized into a corpus of 534,211,366 tokens. Then, he transformed 2,630 English infinitives as base verbal forms to possible anglicisms by using German morphology rules. After some additional checks and depending on if the potential anglicisms existed in the Twitter corpus, the resulting words were added to the anglicism corpus. With this approach, new anglicism data could be created to further expand the anglicism list.

**Sample 9.1:** Segment 37 in "tagesthemen 22:15 Uhr, 18.02.2020" from test set "Anglicisms 2020".

| | |
|---|---|
| **reference** | Mister Bloomberg hat das Recht |
| baseline | Mister Bloomberg hat das rächt |
| comp-crawl | Mr Bloomberg hat das rächt |

### Improving the Evaluation Options

When alternative spellings were used, the benchmark process recorded a word error which weighed as much as if a completely different word was used instead. Sample 9.1 shows the benchmark result of segment 37 in "tagesthemen 22:15 Uhr, 18.02.2020" from test set "Anglicisms 2020" for the baseline and comp-crawl models. While in the reference, the spelling "Mister" was used, comp-crawl recognized the alternative spelling "Mr" which is an abbreviation of the former word. Here, a word error was detected even though the correct word was recognized. This phenomenon also happens with casing (e.g. "GitHub" vs. "Github") and hyphenated spellings (e.g. "Big Picture" vs. "Big-Picture"). For a more realistic interpretation of both the WER and the EER, an option could be implemented in the benchmark process to suppress errors like this.

As mentioned in Section 5.1.3 on page 59, the PER does not take into account the similarity of the supplemented phoneme. Schaden (2006) developed a phonetic distance measure that respects both the edit distance and the phonetic segment similarity of two phoneme sequences. By adding a measure like that, generated anglicism pronunciations could automatically be compared to the canonical pronunciations taken from Wiktionary. This way, the quality of G2P model results could better be evaluated than only using the PER since valid pronunciation variations would result in a lower phonetic distance measure than actual false pronunciations would.

### Adding Pronunciation Variations

Except for the dictionaries created with the Wiktionary approach, all dictionaries only contain exactly one pronunciation for one word. It could be evaluated if generating additional pronunciation variations would improve the anglicism recognition results.

### Using English Common Word Lists as Anglicism Source

An approach that was designed, but could not be executed and evaluated in time was using an English common word list as basis for anglicisms. The assumption behind this approach was that words that are often used in the English language had the potential for being used in the German language as anglicisms. In the discarded approach, three different common word lists were derived from Google's n-gram corpora provided by

the repository gwordlist (hackerb9, 2020), the OpenSubtitles 2018 corpora provided by the repository FrequencyWords (Dave et al., 2020) and the Wiktionary TV and Movie frequency list (Wiktionary, 2006) which was crawled independently. The pronunciaitons can be obtained by using an English pronunciation dictionary like CMUdict. After the English pronunciations have been mapped to their German counterparts using a P2P model, the dictionary is ready to be used as a supplementary anglicism pronunciation dictionary in an ASR model.

#### Generating a full pronunciation dictionary

In this work, all resulting models were evaluated by creating a supplementary anglicism dictionary based on an anglicism list which was added to the baseline ASR model. However, the models resulting from the comparative and MTL approaches are not only able to generate pronunciations for anglicisms, but for native German words as well. Due to their anglicism distinction methods, native German words will be generated differently than anglicisms. In future experiments, the full pronunciation dictionary could be generated by either of the Comparative and MTL models for evaluating the general performance for anglicisms as well as native German words. For better evaluating the MTL approach, a pure Seq2Seq G2P model without anglicism classification task could additionally be created to better judge the influence of the anglicism classification task. Similar to the evaluation in this work, the resulting pronunciation dictionaries could be compared against the monolingual German Sequitur G2P generated baseline pronunciation dictionary by creating ASR models with similar configurations and testing them on various ASR test sets.

# Bibliography


Bavarian Archive for Speech Signals (2013), 'Pronunciation Lexicon PHONOLEX'.
**URL:** *https://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html*

Bayes, T. (1763), *LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S*, Vol. 53 of *Philosophical Transactions*, Royal Society.
**URL:** *https://doi.org/10.1098/rstl.1763.0053*

Bellegarda, J. R. (2005), 'Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy', *Speech Communication* **46**(2), 140–152.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0167639305000336*

Bergstra, J., Bardenet, R., Bengio, Y. and Kégl, B. (2011), Algorithms for Hyper-Parameter Optimization, *in* J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems', Vol. 24, Curran Associates, Inc., pp. 2546–2554.
**URL:** *https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf*

Bisani, M. and Ney, H. (2008), 'Joint-sequence models for grapheme-to-phoneme conversion', *Speech Communication* **50**(5), 434–451.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0167639308000046*

Brownlee, J. (2019), 'How to Choose Loss Functions When Training Deep Learning Neural Networks', Machine Learning Mastery.
**URL:** *https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks*

Brownlee, J. (2020), 'How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification', Machine Learning Mastery.
**URL:** *https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification*

Bruguier, A., Gnanapragasam, D., Johnson, L., Rao, K. and Beaufays, F. (2017), Pronunciation Learning with RNN-Transducers, *in* 'Interspeech 2017', pp. 2556–2560.
**URL:** *http://dx.doi.org/10.21437/Interspeech.2017-47*

Burmasova, S. (2010), 'Empirische Untersuchung der Anglizismen im Deutschen am Material der Zeitung Die WELT (Jahrgänge 1994 und 2004)'.
**URL:** *https://fis.uni-bamberg.de/handle/uniba/227*

Carnegie Mellon University (2014), 'The CMU Pronouncing Dictionary (Version 0.7b)'.
**URL:** *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*

Caruana, R. (1993), Multitask Learning: A Knowledge-Based Source of Inductive Bias, *in* 'Proceedings of the Tenth International Conference on Machine Learning', Morgan Kaufmann, pp. 41–48.
**URL:** *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.3196&rep=rep1&type=pdf*

Caruana, R. (1997), Multitask Learning, PhD thesis, Carnegie Mellon University.
**URL:** *http://reports-archive.adm.cs.cmu.edu/anon/1997/CMU-CS-97-203.pdf*

Chen, S. F. (2003), Conditional and joint models for grapheme-to-phoneme conversion, *in* '8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003', ISCA.
**URL:** *http://www.isca-speech.org/archive/eurospeech_2003/e03_2033.html*

Coats, S. (2019), 'Lexicon geupdated: New German anglicisms in a social media corpus', *European Journal of Applied Linguistics* **7**, 255 – 280.
**URL:** *https://doi.org/10.1515/eujal-2019-0001*

Dave, H., Dascalescu, D. and Lopez, H. (2020), 'FrequencyWords', GitHub repository. Commit 34bb6447a96d4434dc2d4b24e0cf946c523090b3.
**URL:** *https://github.com/hermitdave/FrequencyWords*

Deligne, S. and Bimbot, F. (1997), Inference of variable-length acoustic units for continuous speech recognition, *in* '1997 IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 3, pp. 1731–1734.
**URL:** *https://ieeexplore.ieee.org/document/598858*

Elfers, A. (2020), 'Der Anglizismen-Index 2020 - Deutsch statt Denglisch'.
**URL:** *https://vds-ev.de/denglisch-und-anglizismen/anglizismenindex/ag-anglizismenindex/*

Galescu, L. and Allen, J. F. (2002), Pronunciation of proper names with a joint n-gram

model for bi-directional grapheme-to-phoneme conversion, *in* J. H. L. Hansen and B. L. Pellom, eds, '7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002', ISCA.
**URL:** *http://www.isca-speech.org/archive/icslp_2002/i02_0109.html*

Görlach, M. (2005), *A Dictionary of European Anglicisms: A Usage Dictionary of Anglicisms in Sixteen European Languages*, Oxford University Press.
**URL:** *https://global.oup.com/academic/product/a-dictionary-of-european-anglicisms-9780199283064*

Goronzy, S., Rapp, S. and Kompe, R. (2004), 'Generating non-native pronunciation variants for lexicon adaptation', *Speech Communication* **42**(1), 109–123. Adaptation Methods for Speech Recognition.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0167639303001158*

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y. and He, K. (2017), 'Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour'.
**URL:** *https://research.fb.com/wp-content/uploads/2017/06/imagenet1kin1h5.pdf*

Gref, M., Köhler, J. and Leh, A. (2018), Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research, *in* N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga, eds, 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018', European Language Resources Association (ELRA).
**URL:** *http://www.lrec-conf.org/proceedings/lrec2018/summaries/137.html*

Gref, M., Walter, O., Schmidt, C., Behnke, S. and Köhler, J. (2020), Multi-Staged Cross-Lingual Acoustic Model Adaption for Robust Speech Recognition in Real-World Applications - A Case Study on German Oral History Interviews, *in* 'Proceedings of The 12th Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 6354–6362.
**URL:** *https://www.aclweb.org/anthology/2020.lrec-1.780*

Görlach, M. (1994), 'A Usage Dictionary of Anglicisms in Selected European Languages1', *International Journal of Lexicography* **7**(3), 223–246.
**URL:** *https://doi.org/10.1093/ijl/7.3.223*

hackerb9 (2020), 'gwordlist', GitHub repository. Commit b12a104627a380e75aef82a0727b3bfb5a415f6d.
**URL:** *https://github.com/hackerb9/gwordlist*

Häkkinen, J., Suontausta, J., Riis, S. and Jensen, K. J. (2003), 'Assessing text-to-phoneme mapping strategies in speaker independent isolated word recognition', *Speech Communication* **41**, 455–467.
**URL:** *https://www.sciencedirect.com/science/article/abs/pii/S0167639303000153?via%3Dihub*

He, K., Zhang, X., Ren, S. and Sun, J. (2015), Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *in* 'Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)', ICCV '15, IEEE Computer Society, USA, pp. 1026––1034.
**URL:** *https://doi.org/10.1109/ICCV.2015.123*

Hochreiter, S. and Schmidhuber, J. (1997), 'Long Short-Term Memory', *Neural Computation* **9**(8), 1735–1780.
**URL:** *https://doi.org/10.1162/neco.1997.9.8.1735*

Huang, X., Jin, X., Li, Q. and Zhang, K. (2020), On Construction of the ASR-oriented Indian English Pronunciation Dictionary, *in* 'Proceedings of The 12th Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 6593–6598.
**URL:** *https://www.aclweb.org/anthology/2020.lrec-1.812*

Hunt, J. W. (2019), Anglicisms in German: tsunami or trickle?, *in* A. Koll-Stobbe, ed., 'Informalization and Hybridization of Speech Practices: Polylingual Meaning-Making across Domains, Genres, and Media', Peter Lang, Bern, Schweiz, pp. 25–58.
**URL:** *https://doi.org/10.3726/978-3-653-05414-9*

International Phonetic Association (2015), 'IPA Chart'.
**URL:** *http://www.internationalphoneticassociation.org/content/ipa-chart*

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer Texts in Statistics, Vol. 103, Springer.

Jedicke, P. (2020), 'Neue Wörter im Duden sorgen für Sprach-Kontroverse: DW: 12.08.2020'.

**URL:** *https://www.dw.com/de/neue-wörter-im-duden-sorgen-für-sprach-kontroverse/a-54539751*

Jiang, L., Hon, H. and Huang, X. (1997), Improvements on a trainable letter-to-sound converter, *in* G. Kokkinakis, N. Fakotakis and E. Dermatas, eds, 'EUROSPEECH 1997', ISCA.
**URL:** *http://www.isca-speech.org/archive/eurospeech_1997/e97_0605.html*

Klatte, I. (2020), 'Landtagspräsident: "Software im Landtag hat Sächsisch gelernt"'.
**URL:** *https://www.landtag.sachsen.de/de/service/presse/23332.cshtml*

Kneser, R. and Ney, H. (1995), Improved backing-off for M-gram language modeling, *in* '1995 International Conference on Acoustics, Speech, and Signal Processing', Vol. 1, pp. 181–184.
**URL:** *https://ieeexplore.ieee.org/document/479394*

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. and Talwalkar, A. (2018), 'Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization'.
**URL:** *https://arxiv.org/pdf/1603.06560.pdf*

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B. and Talwalkar, A. (2020a), 'A System for Massively Parallel Hyperparameter Tuning'.
**URL:** *https://arxiv.org/pdf/1810.05934.pdf*

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A. W. and Metze, F. (2020b), Universal phone recognition with a multilingual allophone system, *in* 'ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 8249–8253.
**URL:** *https://arxiv.org/pdf/2002.11800.pdf*

Max Planck Institute for Psycholinguistics, The Language Archive (2020), 'ELAN (Version 5.9) and Simple-ELAN (Version 1.3) [Computer software]'.
**URL:** *https://archive.mpi.nl/tla/elan*

McCann, B., Keskar, N. S., Xiong, C. and Socher, R. (2018), 'The Natural Language Decathlon: Multitask Learning as Question Answering'.
**URL:** *https://arxiv.org/pdf/1806.08730.pdf*

Milde, B. (2019), 'Speech lex edit', GitHub repository. Commit

0abad026eb9afb9fb80f10c48215935601610266.
**URL:** *https://github.com/uhh-lt/speech-lex-edit*

Milde, B., Schmidt, C. and Köhler, J. (2017), Multitask Sequence-to-Sequence Models for Grapheme-to-Phoneme Conversion, *in* 'Interspeech 2017', pp. 2536–2540.
**URL:** *http://dx.doi.org/10.21437/Interspeech.2017-1436*

Miranda, L. J. (2017), 'Understanding softmax and the negative log-likelihood', Lj Miranda.
**URL:** *https://ljvmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-log-likelihood/*

Moore, R. K. and Skidmore, L. (2019), On the Use/Misuse of the Term 'Phoneme', *in* 'Interspeech 2019', pp. 2340–2344.
**URL:** *http://dx.doi.org/10.21437/Interspeech.2019-2711*

Ney, H. and Ortmanns, S. (1999), 'Dynamic programming search for continuous speech recognition', *IEEE Signal Processing Magazine* **16**(5), 64–83.
**URL:** *https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=790984*

Olah, C. (2015), 'Understanding LSTM Networks'.
**URL:** *http://colah.github.io/posts/2015-08-Understanding-LSTMs/*

Optuna (2018), 'optuna.samplers.TPESampler'.
**URL:** *https://optuna.readthedocs.io/en/stable/reference/generated/optuna.samplers.TPESampler.html*

Pascual, S. (2018), 'Toy example in pytorch for binary classification'.
**URL:** *https://gist.github.com/santi-pdp/d0e9002afe74db04aa5bbff6d076e8fe*

Patel, A., Li, D., Cho, E. and Aleksic, P. (2018), Cross-Lingual Phoneme Mapping for Language Robust Contextual Speech Recognition, *in* '2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 5924–5928.
**URL:** *https://ieeexplore.ieee.org/document/8461600*

Phan, D. (2017), 'How to build a Grapheme-to-Phoneme (G2P) model using PyTorch'.
**URL:** *https://fehiepsi.github.io/blog/grapheme-to-phoneme*

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y. and

Khudanpur, S. (2016), Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI, *in* 'Interspeech 2016', pp. 2751–2755.
**URL:** *http://dx.doi.org/10.21437/Interspeech.2016-595*

Reetz, H. and Jongman, A. (2020), *Phonetics: Transcription, Production, Acoustics, and Perception*, Blackwell Textbooks in Linguistics, Wiley.

RP Online (2013), 'Forscher: "Deutsche Sprache ist nicht von Anglizismen bedroht"'.
**URL:** *https://rp-online.de/panorama/wissen/deutsche-sprache-ist-nicht-von-anglizismen-bedroht_aid-14581071*

Ruder, S. (2017), 'An Overview of Multi-Task Learning in Deep Neural Networks'.
**URL:** *https://arxiv.org/pdf/1706.05098.pdf*

Santos, W. (2015), 'Arpabet-to-IPA', GitHub repository. Commit 043a5050d43724194a5734037397279577cddef7.
**URL:** *https://github.com/wwesantos/arpabet-to-ipa*

Schaden, S. (2003), Rule-based lexical modelling of foreign-accented pronunciation variants, *in* '10th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Budapest, Hungary.
**URL:** *https://www.aclweb.org/anthology/E03-1045*

Schaden, S. (2006), Evaluation of Automatically Generated Transcriptions of Non-Native Pronunciations using a Phonetic Distance Measure, *in* 'Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)', European Language Resources Association (ELRA), Genoa, Italy.
**URL:** *http://www.lrec-conf.org/proceedings/lrec2006/pdf/691_pdf.pdf*

Schlippe, T., Ochs, S. and Schultz, T. (2012), Grapheme-to-Phoneme Model Generation for Indo-European Languages, *in* '37th International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan'. ICASSP 2012.
**URL:** *https://www.csl.uni-bremen.de/cms/images/documents/publications/ICASSP2012-Schlippe_G2PModelGenerationIndoEuropean.pdf*

Schlippe, T., Ochs, S. and Schultz, T. (2014), 'Web-based tools and methods for rapid pronunciation dictionary creation', *Speech Communication* **56**, 101–118.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0167639313000885*

Schmidt, C. A. (2020), 'Audio Mining System for the ARD'.
**URL:** *https://www.iais.fraunhofer.de/en/business-areas/speech-technologies/audio-mining-ard.html*

Sokolov, A., Rohlin, T. and Rastrow, A. (2019), Neural Machine Translation for Multilingual Grapheme-to-Phoneme Conversion, *in* 'Interspeech 2019', pp. 2065–2069.
**URL:** *http://dx.doi.org/10.21437/Interspeech.2019-3176*

Stadtschnitzer, M. (2018), Robust Speech Recognition for German and Dialectal Broadcast Programmes, PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.
**URL:** *https://bonndoc.ulb.uni-bonn.de/xmlui/bitstream/handle/20.500.11811/7658/5236.pdf*

Sutskever, I., Vinyals, O. and Le, Q. V. (2014), Sequence to Sequence Learning with Neural Networks, *in* 'Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2', NIPS'14, MIT Press, Cambridge, MA, USA, pp. 3104––3112.
**URL:** *https://arxiv.org/pdf/1409.3215.pdf*

Toshniwal, S. and Livescu, K. (2016), Read, attend and pronounce: An attention-based approach for grapheme-to-phoneme conversion, *in* 'Workshop on Machine Learning in Speech and Language Processing (MLSLP), Interspeech'.
**URL:** *https://ttic.uchicago.edu/∼klivescu/MLSLP2016/toshniwal.pdf*

van Hessen, A., Broekhuizen, M., Scagliola, S., Draxler, C., Karrouche, N., van den Heuvel, H. and Calamai, S. (n.d.), 'G2P', Oral History & Technology.
**URL:** *https://oralhistory.eu/technology/g2p*

Viterbi, A. (1967), 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Transactions on Information Theory* **13**(2), 260–269.
**URL:** *https://ieeexplore.ieee.org/document/1054010*

Wikimedia Commons (2015), 'Long Short-Term Memory'.
**URL:** *https://commons.wikimedia.org/wiki/File:Long_Short_Term_Memory.png*

Wiktionary (2006), 'Frequency lists - TV and movie scripts', Wiktionary.
**URL:** *https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#TV_and_movie_scripts*

Yao, K. and Zweig, G. (2015), 'Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion', *ArXiv* **abs/1506.00196**.
**URL:** *https://arxiv.org/pdf/1506.00196.pdf*

Yu, D. and Deng, L. (2014), *Automatic Speech Recognition: A Deep Learning Approach*, Signals and Communication Technology, Springer London.
**URL:** *https://link.springer.com/book/10.1007/978-1-4471-5779-3*

# Appendix

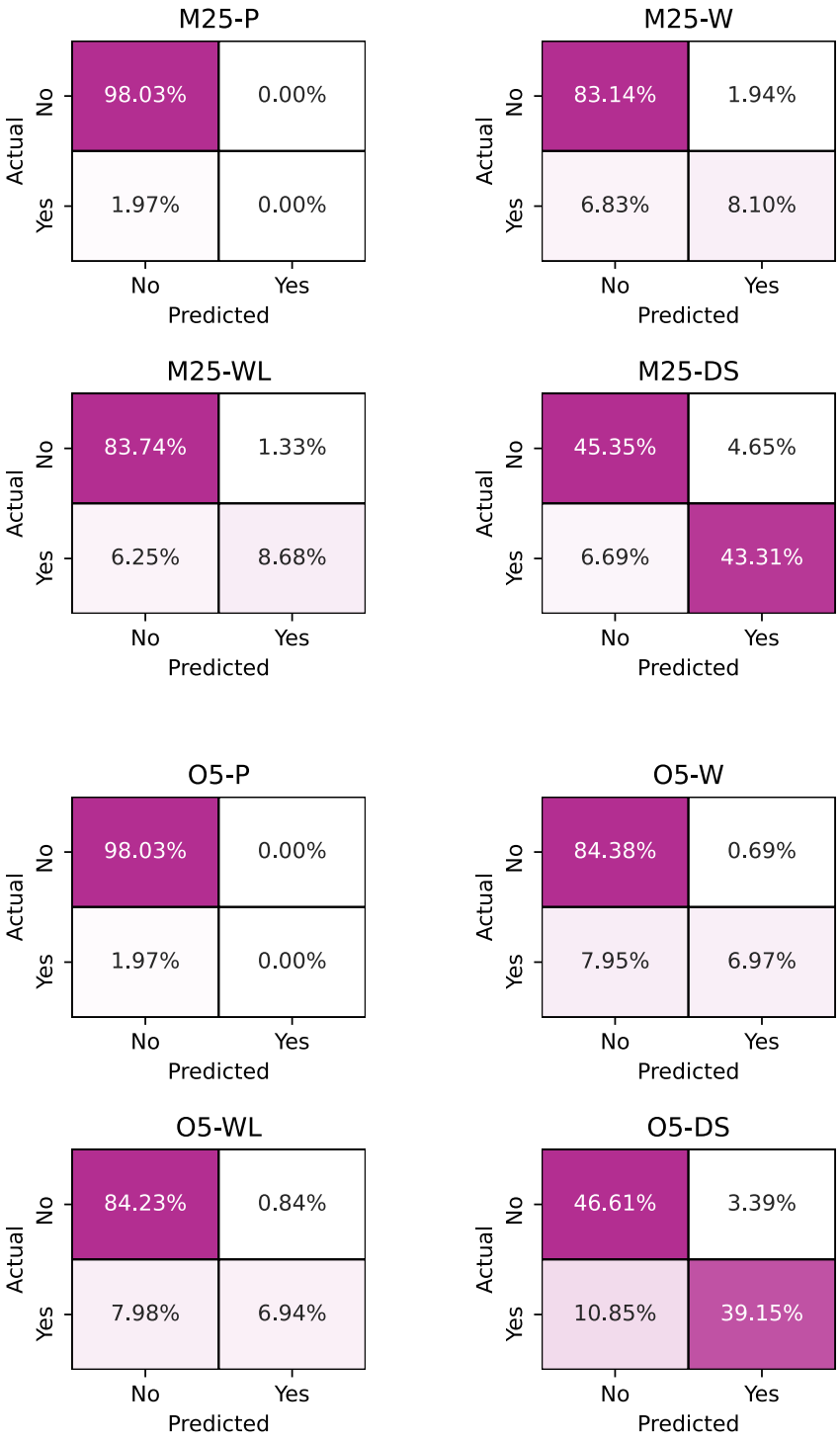


**Figure A1:** Confusion matrices showing the relative anglicism classification results of all MTL models after applying the respective test set. The y-axis shows the actual classification while the x-axis shows the classification predicted by the model. "Yes" and "No" states if the grapheme sequence was classified as an anglicism or not.

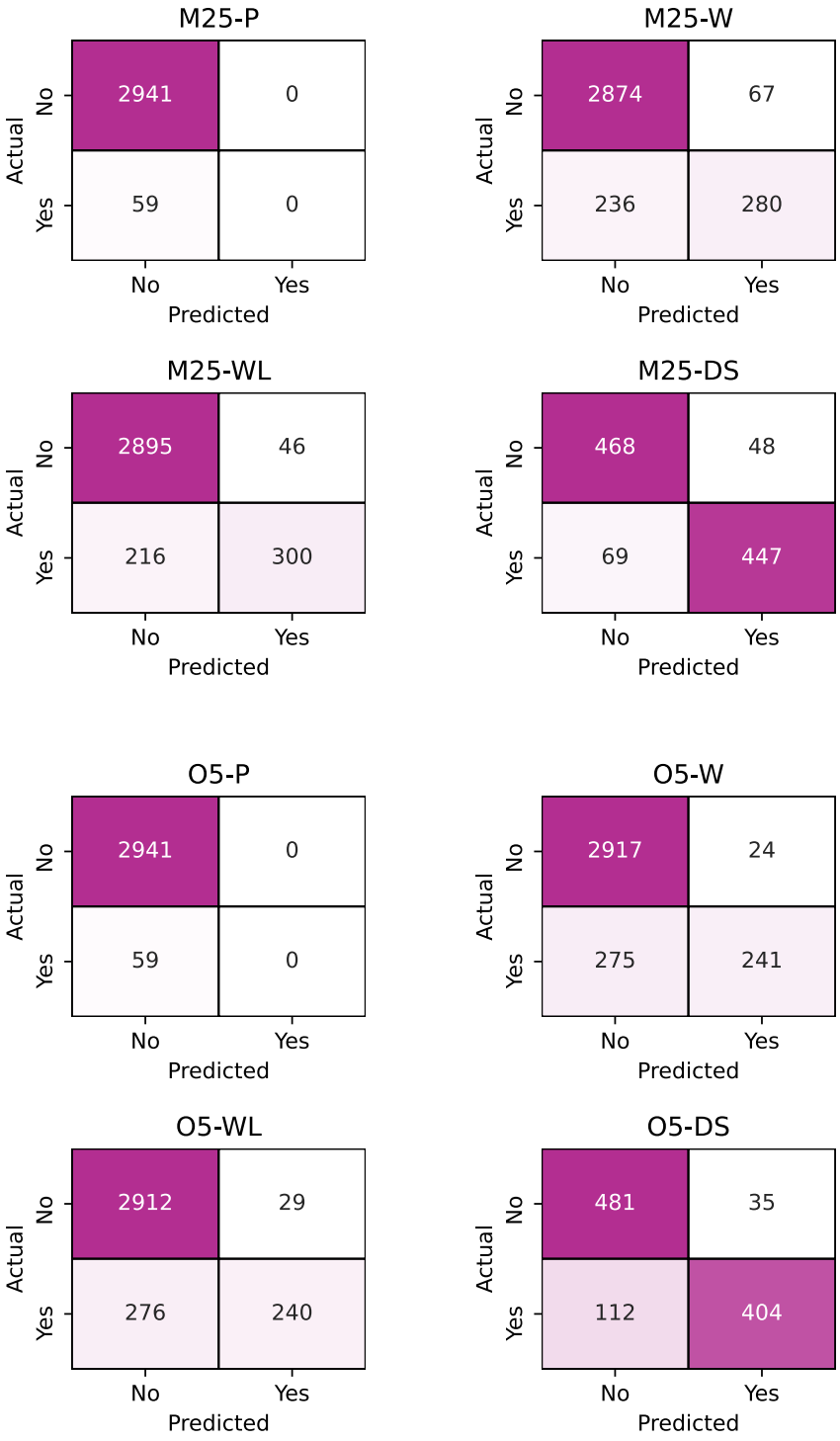


**Figure A2:** Confusion matrices showing the absolute anglicism classification results of all MTL models after applying the respective test set. The y-axis shows the actual classification while the x-axis shows the classification predicted by the model. “Yes” and “No” states if the grapheme sequence was classified as an anglicism or not.

| Name | URL | Context | Duration (hh:mm:ss) |
|---|---|---|---|
| Business Consulting I (Einführung) | https://integration.iubh-dualesstudium.de/kurs/business-consulting-i-einfuehrung/ | Business | 00:14:11 |
| Deutsche Medizintechnik weltweit gefragt | https://youtu.be/0S7ZPB0f1q0 | News segment | 00:04:56 |
| Die Heldenreise des Pre-Sales Consultant | https://youtu.be/ffe5xk9_iLE | Business | 00:03:38 |
| LOGIC Portal Framework | https://youtu.be/T0HUJiZ8l8s | Technical | 00:07:39 |
| Rezo - Die Zerstörung der CDU | https://youtu.be/4Y1lZQsyuSQ | Colloquial | 00:54:57 |
| Rezo - Die Zerstörung der Presse | https://youtu.be/hkncijUZGKA | Colloquial | 00:59:58 |
| Rezo - Wie Politiker momentan auf Schüler scheißen | https://youtu.be/ZiYLQXS-ufs | Colloquial | 00:19:28 |
| Shareholder Value Investment Philosophie | https://youtu.be/7aPMDcoFQac | Business | 00:06:07 |
| tagesschau 20:00 Uhr, 22.01.2020 | https://youtu.be/x_Vi87kMJhM | News segment | 00:15:14 |
| tagesschau 20:00 Uhr, 01.02.2020 | https://youtu.be/nlxN3FGRPDo | News segment | 00:15:28 |
| tagesschau 20:00 Uhr, 03.02.2020 | https://youtu.be/HSMtkz0S34w | News segment | 00:15:30 |
| tagesschau 20:00 Uhr, 11.02.2020 | https://youtu.be/5_VEmc_6z1M | News segment | 00:15:33 |
| tagesthemen 22:15 Uhr, 18.02.2020 | https://youtu.be/lWT1-RiU-EY | News segment | 00:30:19 |
| tagesschau 20:00 Uhr, 21.02.2020 | https://youtu.be/PhG93Y2Vljk | News segment | 00:15:28 |
| tagesschau 20:00 Uhr, 03.03.2020 | https://youtu.be/N1MTOt1up4o | News segment | 00:15:25 |
| nachtmagazin 00:15 Uhr, 04.03.2020 | http://media.tagesschau.de/video/2020/0304/TV-20200304-0049-0200.webxl.h264.mp4 | News segment | 00:19:23 |
| Tagesschau in 100 Sekunden 05:17 Uhr 04.03.2020 | https://media.tagesschau.de/video/100s/2020/0304/TV-100s-0517.webxl.h264.mp4 | News segment | 00:01:51 |
| Venix - Corona-Warn-App | https://youtu.be/x5NcGmYStGI | Technical, colloquial | 00:08:28 |
| Venix - TechNews 94 | https://youtu.be/dxqU6QNRBj0 | Technical, colloquial | 00:03:59 |
| Venix - TechNews 95 | https://youtu.be/7vR4PgQNf1Q | Technical, colloquial | 00:07:12 |
| Venix - TechNews 96 | https://youtu.be/21IneVMf9V8 | Technical, colloquial | 00:05:28 |
| codecentric.AI Bootcamp - Was ist Machine Learning | https://youtu.be/iX9r8wvjKdo | Technical | 00:34:27 |

**Table A1:** Videos used in the testset “Anglicisms 2020”

| Word | wiki-base, wiki-v1& wiki-v2 |
|---|---|
| abcashen | Q a p k E S n |
| babysittet | b e: b i z I t @ t |
| Bookmarks | b U k m a r k s<br>b U k m a:6 k s |
| chartere | S a r t @ r @<br>S a:6 t @ r @<br>tS a r t @ r @<br>tS a:6 t @ r @ |
| Crowdfundings | k r aU d f a n d I N s |
| durchtrainierter | d U6 C t r E n i:6 t 6 |
| flirte | f l 96 t @ |
| geskatet | g @ s k a: t @ t<br>g @ s k E I t @ t<br>g @ s k e: t @ t |
| Hitparade | h I t p a r a: d @ |
| Jokes | dZ o U k s<br>dZ o: k s |
| Likes | l aI k s |
| Mockumentarys | m O k j u m E n t @ r i s |
| Partners | p a r t n 6 s |
| Pullover | p U l o: v 6<br>p U l Q o: v 6 |
| Roadshows | r o U d S o U s<br>r o: tS o: s |
| Skinheads | s k I n h E ts |
| Squashs | s k v O S s |
| Sweaters | s v E t 6 s<br>s v e: t 6 s |
| Trainings | t r E: n I N s<br>t r e: n I N s |
| VIPs | v I p s |

**Table A2:** Examples for the anglicism pronunciation dictionary contents for all Wiktionary models.

| File | Baseline | wiki-base | wiki-v1 | wiki-v2 |
|---|---|---|---|---|
| Business Consulting I (Einführung) | **10.88%** | **10.88%** | **10.88%** | **10.88%** |
| Deutsche Medizintechnik weltweit gefragt | **22.22%** | **22.22%** | **22.22%** | **22.22%** |
| Die Heldenreise des Pre-Sales Consultant | **29.44%** | **29.44%** | **29.44%** | **29.44%** |
| LOGIC Portal Framework | **25.28%** | **25.28%** | **25.28%** | **25.28%** |
| Rezo - Die Zerstörung der CDU | 18.83% | 18.35% | **18.26%** | **18.26%** |
| Rezo - Die Zerstörung der Presse | 15.33% | **15.07%** | **15.07%** | **15.07%** |
| Rezo - Wie Politiker momentan auf Schüler scheißen | 16.67% | **15.67%** | **15.67%** | **15.67%** |
| Shareholder Value Investment Philosophie | **17.38%** | **17.38%** | **17.38%** | **17.38%** |
| tagesschau 20:00 Uhr, 22.01.2020 | **5.80%** | **5.80%** | **5.80%** | **5.80%** |
| tagesschau 20:00 Uhr, 01.02.2020 | **8.60%** | **8.60%** | **8.60%** | **8.60%** |
| tagesschau 20:00 Uhr, 03.02.2020 | **7.36%** | **7.36%** | 7.55% | 7.55% |
| tagesschau 20:00 Uhr, 11.02.2020 | **10.05%** | **10.05%** | **10.05%** | **10.05%** |
| tagesthemen 22:15 Uhr, 18.02.2020 | 8.97% | **8.76%** | 8.97% | **8.76%** |
| tagesschau 20:00 Uhr, 21.02.2020 | **7.94%** | **7.94%** | **7.94%** | **7.94%** |
| tagesschau 20:00 Uhr, 03.03.2020 | **6.67%** | **6.67%** | **6.67%** | **6.67%** |
| nachtmagazin 00:15 Uhr, 04.03.2020 | **12.29%** | **12.29%** | **12.29%** | **12.29%** |
| Tagesschau in 100 Sekunden 05:17 Uhr, 04.03.2020 | **12.20%** | **12.20%** | **12.20%** | **12.20%** |
| Venix - Corona-Warn-App | **14.42%** | 14.61% | 14.61% | 14.61% |
| Venix - TechNews 94 | 15.81% | 15.81% | **15.37%** | **15.37%** |
| Venix - TechNews 95 | **19.97%** | **19.97%** | **19.97%** | **19.97%** |
| Venix - TechNews 96 | **19.86%** | 20.00% | 20.00% | 20.00% |
| codecentric.AI Bootcamp - Was ist Machine Learning | **18.43%** | 18.50% | 18.50% | 18.50% |
| **Total** | 15.80% | 15.72% | 15.71% | **15.70%** |

**Table A3:** WERs per file in test set “Anglicisms 2020” for all Wiktionary models.

| File | Baseline | wiki-base | wiki-v1 | wiki-v2 |
|---|---|---|---|---|
| Inside Bergisch Gladbach - Portrait Ermittlerinnen in Missbrauchsfällen | 9.56% | 9.56% | 9.56% | **9.40%** |
| Streit um die Sommerferien | **5.41%** | **5.41%** | **5.41%** | **5.41%** |
| Richter werten CumEx-Geschäfte als strafbar | **9.09%** | **9.09%** | **9.09%** | **9.09%** |
| Kollegengespraech - Nach dem Nato-Gipfel | **5.94%** | **5.94%** | **5.94%** | **5.94%** |
| Telefoninterview mit Knut Giesler - Bezirksleiter IG Metall NRW | **10.56%** | **10.56%** | **10.56%** | **10.56%** |
| Kollegengespraech - Generalstreik in Frankreich | **6.17%** | **6.17%** | **6.17%** | **6.17%** |
| Die Sicht der anderen Parteien vor dem SPD-Parteitag | **2.42%** | **2.42%** | **2.42%** | **2.42%** |
| Bundesverteidigungsministerin Kramp-Karrenbauer im Kundus | **6.50%** | **6.50%** | **6.50%** | **6.30%** |
| Der Limbecker Platz in Essen | 8.47% | 8.62% | 8.62% | **8.19%** |
| Kollegengespraech - IS-Angriff auf Militärstützpunkt in Mali | **4.31%** | **4.31%** | **4.31%** | **4.31%** |
| Kollegengespräch - Sport | 9.81% | 9.81% | 9.81% | **9.64%** |
| Fracking-Stopp in Großbritannien | **4.93%** | **4.93%** | **4.93%** | **4.93%** |
| Auf der Suche nach dem Sinn - Die Woche der CDU nach der Landtagswahl in Thüringen | **2.10%** | **2.10%** | **2.10%** | **2.10%** |
| Besuch von Bundeskanzlerin Merkel in Indien | **8.47%** | **8.47%** | **8.47%** | **8.47%** |
| UN-Entscheidung - Madrid richtet statt Chile den Weltklimagipfel aus | **3.56%** | **3.56%** | **3.56%** | **3.56%** |
| Weitere Proteste in Chile | **5.46%** | **5.46%** | **5.46%** | **5.46%** |
| Polizeigewalt gegen Demonstranten in Hongkong | **3.74%** | **3.74%** | **3.74%** | **3.74%** |
| Kollegengespraech - Morddrohung gegen die Grünen-Abgeordneten | **3.88%** | **3.88%** | **3.88%** | **3.88%** |
| **Total** | 6.56% | 6.57% | 6.57% | **6.51%** |

**Table A4:** WERs per file in test set “German Broadcast 2020” for all Wiktionary models.

| File | Baseline | wiki-base | wiki-v1 | wiki-v2 |
|---|---|---|---|---|
| Antrittsbesuch in der Heimatstadt | **10.10%** | **10.10%** | **10.10%** | **10.10%** |
| bremenports baut Terminal in Island | **10.61%** | **10.61%** | **10.61%** | **10.61%** |
| Deutscher Schulpreis an Gesamtschule Ost | **9.62%** | **9.62%** | **9.62%** | **9.62%** |
| Die Neue in der Bremer GAK | 14.00% | 14.00% | **13.79%** | 14.00% |
| Fly-Over ab heute gesperrt | **8.97%** | **8.97%** | **8.97%** | **8.97%** |
| Ghostnet Kunst im Übersee-Museum Naue | **8.72%** | 8.90% | 8.90% | 8.90% |
| Musik und Licht am Hollersee | **12.47%** | **12.47%** | **12.47%** | **12.47%** |
| Nachbericht Radio-Bremen-Krimipreisverleihung 2018 | **10.77%** | **10.77%** | **10.77%** | **10.77%** |
| Porträt Carsten Meyer-Heder | **9.92%** | **9.92%** | **9.92%** | **9.92%** |
| Reportage 10 Jahre Waterfront | 11.55% | **11.01%** | **11.01%** | **11.01%** |
| Reportage vom Inneren der Bgm.-Smidt-Brücke | **11.30%** | 11.58% | **11.30%** | 11.58% |
| Schaefer statt Linnert bei den Grünen Bremen | **5.37%** | **5.37%** | **5.37%** | **5.37%** |
| Schauplatz Nordwest - Der Bremer Feigenbaum in der Überseestadt | **11.91%** | **11.91%** | **11.91%** | **11.91%** |
| Skulptour - Mitmachausstellung im Kek-Kindermuseum | **10.75%** | **10.75%** | **10.75%** | **10.75%** |
| swb-Marathon - Zusammenfassung | **16.15%** | **16.15%** | **16.15%** | **16.15%** |
| Wirtschaft regional - Teeherstellung in Bremen | **12.58%** | **12.58%** | **12.58%** | **12.58%** |
| **Total** | 10.84% | 10.83% | **10.80%** | 10.83% |

**Table A5:** WERs per file in test set “Challenging Broadcast 2018” for all Wiktionary models.

| Name | Baseline | wiki-base | wiki-v1 | wiki-v2 |
|---|---|---|---|---|
| Business Consulting I (Einführung) | **12.12%** | **12.12%** | **12.12%** | **12.12%** |
| Deutsche Medizintechnik weltweit gefragt | **50.00%** | **50.00%** | **50.00%** | **50.00%** |
| Die Heldenreise des Pre-Sales Consultant | **34.48%** | 37.93% | 37.93% | 37.93% |
| LOGIC Portal Framework | **61.29%** | **61.29%** | **61.29%** | **61.29%** |
| Rezo - Die Zerstörung der CDU | 44.79% | 41.67% | **40.63%** | **40.63%** |
| Rezo - Die Zerstörung der Presse | 48.51% | **45.96%** | **45.96%** | **45.96%** |
| Rezo - Wie Politiker momentan auf Schüler scheißen | 57.69% | **46.15%** | **46.15%** | **46.15%** |
| Shareholder Value Investment Philosophie | **53.85%** | **53.85%** | **53.85%** | **53.85%** |
| tagesschau 20:00 Uhr, 22.01.2020 | **20.00%** | **20.00%** | **20.00%** | **20.00%** |
| tagesschau 20:00 Uhr, 01.02.2020 | **6.25%** | **6.25%** | **6.25%** | **6.25%** |
| tagesschau 20:00 Uhr, 03.02.2020 | **15.94%** | **15.94%** | **15.94%** | **15.94%** |
| tagesschau 20:00 Uhr, 11.02.2020 | **11.76%** | **11.76%** | **11.76%** | **11.76%** |
| tagesthemen 22:15 Uhr, 18.02.2020 | **20.00%** | **20.00%** | **20.00%** | **20.00%** |
| tagesschau 20:00 Uhr, 21.02.2020 | **12.50%** | **12.50%** | **12.50%** | **12.50%** |
| tagesschau 20:00 Uhr, 03.03.2020 | **20.59%** | **20.59%** | **20.59%** | **20.59%** |
| nachtmagazin 00:15 Uhr, 04.03.2020 | **17.74%** | **17.74%** | **17.74%** | **17.74%** |
| Tagesschau in 100 Sekunden 05:17 Uhr, 04.03.2020 | **25.00%** | **25.00%** | **25.00%** | **25.00%** |
| Venix - Corona Warn App | **26.00%** | **26.00%** | **26.00%** | **26.00%** |
| Venix - TechNews 94 | 36.84% | 36.84% | **35.09%** | **35.09%** |
| Venix - TechNews 95 | **45.31%** | **45.31%** | **45.31%** | **45.31%** |
| Venix - TechNews 96 | **47.14%** | **47.14%** | **47.14%** | **47.14%** |
| codecentric.AI Bootcamp - Was ist Machine Learning | **51.81%** | 52.17% | 52.17% | 52.17% |
| **Total** | 39.50% | 38.77% | **38.62%** | **38.62%** |

**Table A6:** EERs per file in test set “Anglicisms 2020” for all Wiktionary models.

| Word | comp-0 | comp-0.25 | comp-0.5 | comp-0.75 | comp-1 | comp-crawl |
|---|---|---|---|---|---|---|
| abcashen | Q a p k a S @ n | Q a p k E6 S I n | Q a p k E6 S I n | Q a p k E6 S I n | Q E6 p k E6 S I n | *N/A* |
| babysittet | b e: b i s I t @ t | b e: b i s I t @ t | b e: b i s I t @ t | b e: b i s I t @ t | b e: b i s I t @ t | *N/A* |
| Bookmarks | b U k m a r k s | b U k m a r k s | b U6 k m a r k s | b U6 k m a r k s | b U6 k m a r k s | *N/A* |
| chartere | tS a r t 6 | S a r t 6 | tS a r t 6 | tS a: l t @ | tS a: l t @ | *N/A* |
| Crowdfundings | k r u: t f U n d I N s | k r a U t f U n d I N s | k r o U t f U n d I N s | k 96 f a n d E N s | k r @ U t f a n d E N s | k r u: t f U n d I N s |
| durchtrainierter | d U6 C t r e n i:6 t 6 | d U6 C t r e n i:6 t 6 | d U6 C t r e n i:6 t 6 | d U6 C t r e n i:6 t 6 | d U6 C t r e n i:6 t 6 | *N/A* |
| flirte | f l 96 t @ | f l 96 t @ | f l 96 t @ | f l 96 t @ | f l 96 t @ | *N/A* |
| geskatet | g @ s k e: t @ t | g @ s k e: t @ t | g @ s k e: t @ t | g @ s k e: t @ t | g @ s k e: t @ t | *N/A* |
| Hitparade | h I t p a r a: d @ | h I t p a r a: d @ | h I t p a r a: d @ | h I t p a r a: d @ | h I t p a r a: d @ | *N/A* |
| Jokes | j o: k @ s | j o: k @ s | j o: k s | dZ E U k s | dZ E U k s | *N/A* |
| Likes | l i k @ s | l aI k s | l aI k s | l aI k s | l aI k s | l aI k s |
| Mockumentarys | m O k u m e: n @ r i s | m a k j m E n @ r i s | m a: k j u: m e n @ r i s | m a: k j u: m E n O r i s | m a: k j u: m E n @ r i s | m a: k j m E n @ r i s |
| Partners | p a r t n 6 s | p a r t n 6 s | p a r t n 6 s | p a r t n 6 s | p a r t n 6 s | *N/A* |
| Pullover | p U l o: v 6 | p U l o: v 6 | p U l o: v 6 | p U l o: v 6 | p U l o: v 6 | *N/A* |
| Roadshows | r o: t S o: s | r o: t S o: s | r o: t S o: s | r o: t S o: s | r o: t S o: s | *N/A* |
| Skinheads | S i: n h e a: ts | S i: n h e a: ts | S k I n h e a: ts | S i: n h e a: ts | z k I6 n h E l s | S i: n h e a: ts |
| Squashs | s k v a S @ s | s k v a s C s | k O6 S @ s | k u a S s | k u a S s | *N/A* |
| Sweaters | s v E t 6 s | s v E t 6 s | s v E t 6 s | ts v E d 2: s | ts v E d 2: s | *N/A* |
| Trainings | t r e: n I N s | t r e: n I N s | t r e: n I N s | t r e: n I N s | t r e: n I N s | *N/A* |
| VIPs | v I p s | v e: a e: s | v e: a e: s | v e: a e: s | v e: a e: s | *N/A* |

**Table A7:** Examples for the anglicism pronunciation dictionary contents of all Comparative models.

| File | Baseline | comp-0 | comp-0.25 | comp-0.5 | comp-0.75 | comp-1 | comp-crawl |
|---|---|---|---|---|---|---|---|
| Business Consulting I (Einführung) | **10.88%** | **10.88%** | **10.88%** | **10.88%** | **10.88%** | **10.88%** | **10.88%** |
| Deutsche Medizintechnik weltweit gefragt | **22.22%** | **22.22%** | **22.22%** | **22.22%** | **22.22%** | **22.22%** | **22.22%** |
| Die Heldenreise des Pre-Sales Consultant | 29.44% | 29.44% | 29.44% | 29.44% | 29.44% | 29.44% | **27.57%** |
| LOGIC Portal Framework | 25.28% | 25.28% | 25.28% | 25.28% | **25.12%** | **25.12%** | **25.12%** |
| Rezo - Die Zerstörung der CDU | 18.83% | 18.73% | 18.73% | 18.73% | 18.83% | 18.83% | **18.64%** |
| Rezo - Die Zerstörung der Presse | 15.33% | 15.33% | 15.15% | **15.00%** | **15.00%** | **15.00%** | 15.19% |
| Rezo - Wie Politiker momentan auf Schüler scheißen | 16.67% | 16.33% | 16.00% | **15.00%** | 16.00% | 16.00% | 16.00% |
| Shareholder Value Investment Philosophie | 17.38% | 18.44% | 15.60% | 16.67% | **14.89%** | **14.89%** | 16.67% |
| tagesschau 20:00 Uhr, 22.01.2020 | **5.80%** | **5.80%** | **5.80%** | **5.80%** | **5.80%** | **5.80%** | **5.80%** |
| tagesschau 20:00 Uhr, 01.02.2020 | 8.60% | **8.11%** | **8.11%** | **8.11%** | 8.60% | 8.60% | 8.60% |
| tagesschau 20:00 Uhr, 03.02.2020 | **7.36%** | **7.36%** | **7.36%** | 7.55% | 7.55% | 7.55% | 7.55% |
| tagesschau 20:00 Uhr, 11.02.2020 | **10.05%** | **10.05%** | **10.05%** | **10.05%** | **10.05%** | **10.05%** | **10.05%** |
| tagesthemen 22:15 Uhr, 18.02.2020 | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** | 10.04% |
| tagesschau 20:00 Uhr, 21.02.2020 | **7.94%** | **7.94%** | **7.94%** | **7.94%** | **7.94%** | **7.94%** | **7.94%** |
| tagesschau 20:00 Uhr, 03.03.2020 | **6.67%** | **6.67%** | **6.67%** | **6.67%** | **6.67%** | **6.67%** | 7.18% |
| nachtmagazin 00:15 Uhr, 04.03.2020 | **12.29%** | **12.29%** | **12.29%** | **12.29%** | **12.29%** | **12.29%** | 13.04% |
| Tagesschau in 100 Sekunden 05:17 Uhr, 04.03.2020 | **12.20%** | **12.20%** | **12.20%** | **12.20%** | **12.20%** | **12.20%** | **12.20%** |
| Venix - Corona-Warn-App | **14.42%** | 14.51% | 14.51% | **14.42%** | 14.70% | 14.70% | 14.80% |
| Venix - TechNews 94 | **15.81%** | 16.26% | 16.04% | 16.04% | 16.04% | 16.04% | 16.26% |
| Venix - TechNews 95 | **19.97%** | **19.97%** | **19.97%** | **19.97%** | 20.14% | 20.14% | 20.14% |
| Venix - TechNews 96 | 19.86% | **19.44%** | **19.44%** | 19.58% | 19.58% | 19.58% | 20.56% |
| codecentric.AI Bootcamp - Was ist Machine Learning | 18.43% | 18.46% | 18.38% | **18.06%** | 18.30% | 18.30% | 18.58% |
| **Total** | 15.80% | 15.80% | 15.67% | **15.59%** | 15.67% | 15.67% | 15.89% |

**Table A8:** WERs per file in test set “Anglicisms 2020” for all Comparative models.

| File | Baseline | comp-0 | comp-0.25 | comp-0.5 | comp-0.75 | comp-1 | comp-crawl |
|---|---|---|---|---|---|---|---|
| Inside Bergisch Gladbach - Portrait Ermittlerinnen in Missbrauchsfällen | **9.56%** | **9.56%** | **9.56%** | **9.56%** | **9.56%** | **9.56%** | **9.56%** |
| Streit um die Sommerferien | **5.41%** | **5.41%** | **5.41%** | **5.41%** | **5.41%** | **5.41%** | **5.41%** |
| Richter werten CumEx-Geschäfte als strafbar | **9.09%** | **9.09%** | **9.09%** | **9.09%** | **9.09%** | **9.09%** | **9.09%** |
| Kollegengespraech - Nach dem Nato-Gipfel | **5.94%** | **5.94%** | **5.94%** | **5.94%** | **5.94%** | **5.94%** | **5.94%** |
| Telefoninterview mit Knut Giesler - Bezirksleiter IG Metall NRW | **10.56%** | **10.56%** | **10.56%** | **10.56%** | **10.56%** | **10.56%** | **10.56%** |
| Kollegengespraech - Generalstreik in Frankreich | **6.17%** | **6.17%** | **6.17%** | **6.17%** | **6.17%** | **6.17%** | **6.17%** |
| Die Sicht der anderen Parteien vor dem SPD-Parteitag | **2.42%** | **2.42%** | **2.42%** | **2.42%** | **2.42%** | **2.42%** | **2.42%** |
| Bundesverteidigungsministerin Kramp-Karrenbauer im Kundus | **6.50%** | **6.50%** | **6.50%** | **6.50%** | **6.50%** | **6.50%** | **6.50%** |
| Der Limbecker Platz in Essen | **8.47%** | 8.62% | 8.62% | 8.62% | 8.62% | 8.62% | 8.62% |
| Kollegengespraech - IS-Angriff auf Militärstützpunkt in Mali | **4.31%** | **4.31%** | **4.31%** | **4.31%** | **4.31%** | **4.31%** | **4.31%** |
| Kollegengespräch - Sport | **9.81%** | **9.81%** | **9.81%** | **9.81%** | **9.81%** | **9.81%** | **9.81%** |
| Fracking-Stopp in Großbritannien | **4.93%** | **4.93%** | **4.93%** | **4.93%** | **4.93%** | **4.93%** | **4.93%** |
| Auf der Suche nach dem Sinn - Die Woche der CDU nach der Landtagswahl in Thüringen | **2.10%** | **2.10%** | **2.10%** | **2.10%** | **2.10%** | **2.10%** | **2.10%** |
| Besuch von Bundeskanzlerin Merkel in Indien | **8.47%** | **8.47%** | **8.47%** | **8.47%** | **8.47%** | **8.47%** | **8.47%** |
| UN-Entscheidung - Madrid richtet statt Chile den Weltklimagipfel aus | **3.56%** | **3.56%** | **3.56%** | **3.56%** | **3.56%** | **3.56%** | **3.56%** |
| Weitere Proteste in Chile | **5.46%** | **5.46%** | **5.46%** | **5.46%** | **5.46%** | **5.46%** | **5.46%** |
| Polizeigewalt gegen Demonstranten in Hongkong | 3.74% | 3.74% | 3.74% | 3.74% | 3.74% | 3.74% | **3.06%** |
| Kollegengespraech - Morddrohung gegen die Grünen-Abgeordneten | **3.88%** | **3.88%** | **3.88%** | **3.88%** | **3.88%** | **3.88%** | **3.88%** |
| **Total** | 6.56% | 6.57% | 6.57% | 6.57% | 6.57% | 6.57% | **6.55%** |

**Table A9:** WERs per file in test set "German Broadcast 2020" for all Comparative models.

| File | Baseline | comp-0 | comp-0.25 | comp-0.5 | comp-0.75 | comp-1 | comp-crawl |
|---|---|---|---|---|---|---|---|
| Antrittsbesuch in der Heimatstadt | **10.10%** | **10.10%** | **10.10%** | **10.10%** | **10.10%** | **10.10%** | 10.28% |
| bremenports baut Terminal in Island | **10.61%** | 10.73% | **10.61%** | **10.61%** | **10.61%** | **10.61%** | **10.61%** |
| Deutscher Schulpreis an Gesamtschule Ost | **9.62%** | **9.62%** | **9.62%** | **9.62%** | **9.62%** | **9.62%** | **9.62%** |
| Die Neue in der Bremer GAK | **14.00%** | **14.00%** | **14.00%** | **14.00%** | **14.00%** | **14.00%** | **14.00%** |
| Fly-Over ab heute gesperrt | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** |
| Ghostnet Kunst im Übersee-Museum Naue | 8.72% | 8.72% | 8.72% | 8.72% | 8.72% | 8.72% | **8.54%** |
| Musik und Licht am Hollersee | **12.47%** | **12.47%** | **12.47%** | **12.47%** | **12.47%** | **12.47%** | **12.47%** |
| Nachbericht Radio-Bremen-Krimipreisverleihung 2018 | 10.77% | 10.77% | 10.77% | 10.77% | 10.77% | 10.77% | **10.57%** |
| Porträt Carsten Meyer-Heder | **9.92%** | **9.92%** | **9.92%** | **9.92%** | **9.92%** | **9.92%** | **9.92%** |
| Reportage 10 Jahre Waterfront | 11.55% | 11.55% | **11.01%** | **11.01%** | **11.01%** | **11.01%** | **11.01%** |
| Reportage vom Inneren der Bgm.-Smidt-Brücke | **11.30%** | **11.30%** | **11.30%** | **11.30%** | **11.30%** | **11.30%** | **11.30%** |
| Schaefer statt Linnert bei den Grünen Bremen | **5.37%** | **5.37%** | **5.37%** | **5.37%** | **5.37%** | **5.37%** | **5.37%** |
| Schauplatz Nordwest - Der Bremer Feigenbaum in der Überseestadt | **11.91%** | **11.91%** | **11.91%** | **11.91%** | **11.91%** | **11.91%** | **11.91%** |
| Skulptour - Mitmachausstellung im Kek-Kindermuseum | **10.75%** | **10.75%** | **10.75%** | **10.75%** | **10.75%** | **10.75%** | **10.75%** |
| swb-Marathon - Zusammenfassung | **16.15%** | 16.43% | 16.43% | 16.43% | 16.43% | 16.43% | 16.43% |
| Wirtschaft regional - Teeherstellung in Bremen | **12.58%** | **12.58%** | **12.58%** | **12.58%** | 12.70% | 12.70% | 12.70% |
| **Total** | 10.84% | 10.86% | **10.82%** | **10.82%** | 10.83% | 10.83% | 10.83% |

**Table A10:** WERs per file in test set “Challenging Broadcast 2018” for all Comparative models.

| Name | Baseline | comp-0 | comp-0.25 | comp-0.5 | comp-0.75 | comp-1 | comp-crawl |
|---|---|---|---|---|---|---|---|
| Business Consulting I (Einführung) | **12.12%** | **12.12%** | **12.12%** | **12.12%** | **12.12%** | **12.12%** | **12.12%** |
| Deutsche Medizintechnik weltweit gefragt | **50.00%** | **50.00%** | **50.00%** | **50.00%** | **50.00%** | **50.00%** | **50.00%** |
| Die Heldenreise des Pre-Sales Consultant | 34.48% | 37.93% | 37.93% | 37.93% | 37.93% | 37.93% | **31.03%** |
| LOGIC Portal Framework | 61.29% | 61.29% | 61.29% | 61.29% | 61.29% | 61.29% | **58.06%** |
| Rezo - Die Zerstörung der CDU | 44.79% | 44.79% | **42.71%** | **42.71%** | 43.75% | 43.75% | **42.71%** |
| Rezo - Die Zerstörung der Presse | 48.51% | 48.09% | 47.23% | **45.96%** | **45.96%** | **45.96%** | 47.66% |
| Rezo - Wie Politiker momentan auf Schüler scheißen | 57.69% | 53.85% | 50.00% | **42.31%** | 50.00% | 50.00% | 50.00% |
| Shareholder Value Investment Philosophie | 53.85% | 53.85% | 50.00% | 53.85% | **46.15%** | **46.15%** | 57.69% |
| tagesschau 20:00 Uhr, 22.01.2020 | **20.00%** | **20.00%** | **20.00%** | **20.00%** | **20.00%** | **20.00%** | **20.00%** |
| tagesschau 20:00 Uhr, 01.02.2020 | 6.25% | **3.13%** | **3.13%** | **3.13%** | 6.25% | 6.25% | 6.25% |
| tagesschau 20:00 Uhr, 03.02.2020 | **15.94%** | **15.94%** | **15.94%** | **15.94%** | **15.94%** | **15.94%** | **15.94%** |
| tagesschau 20:00 Uhr, 11.02.2020 | **11.76%** | **11.76%** | **11.76%** | **11.76%** | **11.76%** | **11.76%** | **11.76%** |
| tagesthemen 22:15 Uhr, 18.02.2020 | **20.00%** | **20.00%** | **20.00%** | **20.00%** | **20.00%** | **20.00%** | 28.57% |
| tagesschau 20:00 Uhr, 21.02.2020 | **12.50%** | **12.50%** | **12.50%** | **12.50%** | **12.50%** | **12.50%** | **12.50%** |
| tagesschau 20:00 Uhr, 03.03.2020 | **20.59%** | **20.59%** | **20.59%** | **20.59%** | **20.59%** | **20.59%** | 23.53% |
| nachtmagazin 00:15 Uhr, 04.03.2020 | **17.74%** | **17.74%** | **17.74%** | **17.74%** | **17.74%** | **17.74%** | 20.97% |
| Tagesschau in 100 Sekunden 05:17 Uhr, 04.03.2020 | **25.00%** | **25.00%** | **25.00%** | **25.00%** | **25.00%** | **25.00%** | **25.00%** |
| Venix - Corona Warn App | 26.00% | 25.00% | 25.00% | **24.00%** | 26.00% | 26.00% | 27.00% |
| Venix - TechNews 94 | 36.84% | 36.84% | **35.09%** | **35.09%** | **35.09%** | **35.09%** | 38.60% |
| Venix - TechNews 95 | 45.31% | 45.31% | 45.31% | 45.31% | 45.31% | 45.31% | **42.19%** |
| Venix - TechNews 96 | 47.14% | **42.86%** | **42.86%** | 44.29% | 44.29% | 44.29% | 48.57% |
| codecentric.AI Bootcamp - Was ist Machine Learning | 51.81% | 52.17% | 51.81% | **50.00%** | 51.09% | 51.09% | 53.26% |
| **Total** | 39.50% | 39.13% | 38.55% | **37.89%** | 38.40% | 38.40% | 39.72% |

**Table A11:** EERs per file in test set “Anglicisms 2020” for all Comparative models.

| Word | mtl-m25-ds | mtl-o5-ds | mtl-m25-wl | mtl-o5-wl |
|---|---|---|---|---|
| abcashen | Q a p k E S n | Q a p k E S n | Q a p k E S n | Q a p k E S n |
| babysittet | b e: b i z I t @ t | b e: b i z I t @ t | b e: b i z I t @ t | b e: b i z I t @ t |
| Bookmarks | b U k m a r k s | b U k m a r k s | b U k m a r k s | b U k m a r k s |
| chartere | tS a r t @ r @ | tS a r t @ r @ | S a r t @ r @ | S a r t @ r @ |
| Crowdfundings | k r aU d f U n d I N s | k r aU d f a n d I N s | k r o: t f U n d I N s | k r aU t f U n d I N s |
| durchtrainierter | d U6 C t r E n i:6 t 6 | d U6 C t r e n i:6 t 6 | d U6 C t r e n i:6 t 6 | d U6 C t r e n i:6 t 6 |
| flirte | f l 96 t @ | f l 96 t @ | f l I6 t @ | f l I6 t @ |
| geskatet | g @ s k E I t @ t | g @ s k E I t @ t | g @ s k E I t @ t | g @ s k E I t @ t |
| Hitparade | h I t p a r a: d @ | h I t p a r a: d @ | h I t p a r a: d @ | h I t p a r a: d @ |
| Jokes | j o: k @ s | j o: k s | dZ O k s | dZ o: k s |
| Likes | l aI k s | l aI k s | l aI k s | l aI k s |
| Mockumentarys | m O k u m E n t @ r i s | m O k u m E n t @ r i s | m O k u m E n t @ r i s | m O k u m E n t @ r i s |
| Partners | p a r t n 6 s | p a r t n 6 s | p a r t n 6 s | p a r t n 6 s |
| Pullover | p U l o: v 6 | p U l Q o: v 6 | p U l Q o: v 6 | p U l o: v 6 |
| Roadshows | r O U d S O U s | r O U d S O U s | r o: tS o: s | r O ts h o: s |
| Skinheads | s k I n h E ts | s k I n h E ts | s k I n h E ts | s k i: n h E ts |
| Squashs | s k v E S s | s k v O S s | s k v E S s | s k v E S s |
| Sweaters | s v e: t 6 s | s v E t 6 s | s v e: t 6 s | s v e: t 6 s |
| Trainings | t r e: n I N s | t r e: n I N s | t r e: n I N s | t r e: n I N s |
| VIPs | v i: p s | v I p s | f aU p s | f aU Q i: p s |

**Table A12:** Examples for the anglicism pronunciation dictionary contents of all MTL models.

| File | Baseline | mtl-m25-ds | mtl-o5-ds | mtl-m25-wl | mtl-o5-wl |
|---|---|---|---|---|---|
| Business Consulting I (Einführung) | **10.88%** | **10.88%** | **10.88%** | **10.88%** | **10.88%** |
| Deutsche Medizintechnik weltweit gefragt | **22.22%** | **22.22%** | **22.22%** | **22.22%** | **22.22%** |
| Die Heldenreise des Pre-Sales Consultant | 29.44% | **28.50%** | 29.44% | 29.44% | 29.44% |
| LOGIC Portal Framework | 25.28% | 25.28% | 25.28% | **24.96%** | 25.12% |
| Rezo - Die Zerstörung der CDU | 18.83% | 18.26% | **18.45%** | **18.16%** | **18.16%** |
| Rezo - Die Zerstörung der Presse | 15.33% | **14.78%** | 15.04% | 14.96% | 15.19% |
| Rezo - Wie Politiker momentan auf Schüler scheißen | 16.67% | 16.00% | **14.67%** | 15.33% | 15.33% |
| Shareholder Value Investment Philosophie | **17.38%** | **17.38%** | **17.38%** | **17.38%** | **17.38%** |
| tagesschau 20:00 Uhr, 22.01.2020 | 5.80% | 5.36% | **4.91%** | 5.36% | 5.36% |
| tagesschau 20:00 Uhr, 01.02.2020 | 8.60% | 8.60% | 8.60% | **8.11%** | 8.60% |
| tagesschau 20:00 Uhr, 03.02.2020 | **7.36%** | 7.55% | 7.55% | 7.55% | 7.55% |
| tagesschau 20:00 Uhr, 11.02.2020 | **10.05%** | **10.05%** | 11.06% | **10.05%** | **10.05%** |
| tagesthemen 22:15 Uhr, 18.02.2020 | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** |
| tagesschau 20:00 Uhr, 21.02.2020 | 7.94% | **6.35%** | 7.94% | 7.94% | 7.94% |
| tagesschau 20:00 Uhr, 03.03.2020 | **6.67%** | **6.67%** | **6.67%** | **6.67%** | **6.67%** |
| nachtmagazin 00:15 Uhr, 04.03.2020 | **12.29%** | **12.29%** | **12.29%** | **12.29%** | **12.29%** |
| Tagesschau in 100 Sekunden 05:17 Uhr, 04.03.2020 | **12.20%** | **12.20%** | **12.20%** | **12.20%** | **12.20%** |
| Venix - Corona-Warn-App | **14.42%** | 14.61% | 14.61% | 14.51% | **14.42%** |
| Venix - TechNews 94 | **15.81%** | **15.81%** | 16.04% | **15.81%** | **15.81%** |
| Venix - TechNews 95 | **19.97%** | 20.65% | 20.31% | 20.31% | 20.65% |
| Venix - TechNews 96 | **19.86%** | 20.28% | 20.14% | 20.14% | 20.28% |
| codecentric.AI Bootcamp - Was ist Machine Learning | 18.43% | 18.42% | 18.42% | **18.38%** | 18.42% |
| **Total** | 15.80% | 15.67% | 15.73% | **15.65%** | 15.73% |

**Table A13:** WERs per file in test set “Anglicisms 2020” for all MTL models.

| File | Baseline | mtl-m25-ds | mtl-o5-ds | mtl-m25-wl | mtl-o5-wl |
|---|---|---|---|---|---|
| Inside Bergisch Gladbach - Portrait Ermittlerinnen in Missbrauchsfällen | **9.56%** | **9.56%** | **9.56%** | **9.56%** | **9.56%** |
| Streit um die Sommerferien | **5.41%** | **5.41%** | **5.41%** | **5.41%** | **5.41%** |
| Richter werten CumEx-Geschäfte als strafbar | **9.09%** | 9.34% | 9.34% | **9.09%** | **9.09%** |
| Kollegengespraech - Nach dem Nato-Gipfel | **5.94%** | **5.94%** | **5.94%** | **5.94%** | **5.94%** |
| Telefoninterview mit Knut Giesler - Bezirksleiter IG Metall NRW | **10.56%** | **10.56%** | **10.56%** | **10.56%** | **10.56%** |
| Kollegengespraech - Generalstreik in Frankreich | **6.17%** | **6.17%** | **6.17%** | **6.17%** | **6.17%** |
| Die Sicht der anderen Parteien vor dem SPD-Parteitag | **2.42%** | **2.42%** | **2.42%** | **2.42%** | **2.42%** |
| Bundesverteidigungsministerin Kramp-Karrenbauer im Kundus | **6.50%** | **6.50%** | **6.50%** | **6.50%** | **6.50%** |
| Der Limbecker Platz in Essen | **8.47%** | 8.62% | 8.62% | **8.47%** | 8.62% |
| Kollegengespraech - IS-Angriff auf Militärstützpunkt in Mali | **4.31%** | **4.31%** | **4.31%** | **4.31%** | **4.31%** |
| Kollegengespräch - Sport | **9.81%** | **9.81%** | **9.81%** | **9.81%** | **9.81%** |
| Fracking-Stopp in Großbritannien | **4.93%** | **4.93%** | **4.93%** | **4.93%** | **4.93%** |
| Auf der Suche nach dem Sinn - Die Woche der CDU nach der Landtagswahl in Thüringen | **2.10%** | **2.10%** | **2.10%** | **2.10%** | **2.10%** |
| Besuch von Bundeskanzlerin Merkel in Indien | **8.47%** | 8.70% | 8.70% | 8.70% | 8.70% |
| UN-Entscheidung - Madrid richtet statt Chile den Weltklimagipfel aus | **3.56%** | **3.56%** | **3.56%** | **3.56%** | **3.56%** |
| Weitere Proteste in Chile | **5.46%** | **5.46%** | **5.46%** | **5.46%** | **5.46%** |
| Polizeigewalt gegen Demonstranten in Hongkong | **3.74%** | **3.74%** | **3.74%** | **3.74%** | **3.74%** |
| Kollegengespraech - Morddrohung gegen die Grünen-Abgeordneten | **3.88%** | **3.88%** | **3.88%** | **3.88%** | **3.88%** |
| **Total** | **6.56%** | 6.60% | 6.60% | 6.57% | 6.59% |

**Table A14:** WERs per file in test set “German Broadcast 2020” for all MTL models.

| File | Baseline | mtl-m25-ds | mtl-o5-ds | mtl-m25-wl | mtl-o5-wl |
|---|---|---|---|---|---|
| Antrittsbesuch in der Heimatstadt | **10.10%** | **10.10%** | **10.10%** | **10.10%** | **10.10%** |
| bremenports baut Terminal in Island | **10.61%** | **10.61%** | **10.61%** | **10.61%** | **10.61%** |
| Deutscher Schulpreis an Gesamtschule Ost | **9.62%** | **9.62%** | **9.62%** | **9.62%** | **9.62%** |
| Die Neue in der Bremer GAK | **14.00%** | **14.00%** | **14.00%** | **14.00%** | **14.00%** |
| Fly-Over ab heute gesperrt | **8.97%** | **8.97%** | **8.97%** | **8.97%** | **8.97%** |
| Ghostnet Kunst im Übersee-Museum Naue | **8.72%** | **8.72%** | **8.72%** | **8.72%** | **8.72%** |
| Musik und Licht am Hollersee | **12.47%** | **12.47%** | **12.47%** | **12.47%** | **12.47%** |
| Nachbericht Radio-Bremen-Krimipreisverleihung 2018 | **10.77%** | **10.77%** | **10.77%** | **10.77%** | **10.77%** |
| Porträt Carsten Meyer-Heder | **9.92%** | **9.92%** | **9.92%** | **9.92%** | **9.92%** |
| Reportage 10 Jahre Waterfront | **11.55%** | **11.55%** | **11.55%** | 11.73% | **11.55%** |
| Reportage vom Inneren der Bgm.-Smidt-Brücke | **11.30%** | **11.30%** | **11.30%** | **11.30%** | **11.30%** |
| Schaefer statt Linnert bei den Grünen Bremen | **5.37%** | **5.37%** | **5.37%** | **5.37%** | **5.37%** |
| Schauplatz Nordwest - Der Bremer Feigenbaum in der Überseestadt | **11.91%** | **11.91%** | **11.91%** | **11.91%** | 12.13% |
| Skulptour - Mitmachausstellung im Kek-Kindermuseum | **10.75%** | 10.93% | 10.93% | **10.75%** | **10.75%** |
| swb-Marathon - Zusammenfassung | **16.15%** | 16.43% | 16.43% | 16.43% | 16.43% |
| Wirtschaft regional - Teeherstellung in Bremen | **12.58%** | 13.08% | 13.08% | **12.58%** | **12.58%** |
| **Total** | **10.84%** | 10.90% | 10.90% | 10.86% | 10.86% |

**Table A15:** WERs per file in test set “Challenging Broadcast 2018” for all MTL models.

| Name | Baseline | mtl-m25-ds | mtl-o5-ds | mtl-m25-wl | mtl-o5-wl |
|---|---|---|---|---|---|
| Business Consulting I (Einführung) | **12.12%** | **12.12%** | **12.12%** | **12.12%** | **12.12%** |
| Deutsche Medizintechnik weltweit gefragt | **50.00%** | **50.00%** | **50.00%** | **50.00%** | **50.00%** |
| Die Heldenreise des Pre-Sales Consultant | **34.48%** | 37.93% | 37.93% | 37.93% | 37.93% |
| LOGIC Portal Framework | 61.29% | 61.29% | 61.29% | **59.68%** | **59.68%** |
| Rezo - Die Zerstörung der CDU | 44.79% | **41.67%** | 43.75% | 43.75% | 43.75% |
| Rezo - Die Zerstörung der Presse | 48.51% | **43.83%** | 45.11% | 45.11% | 46.81% |
| Rezo - Wie Politiker momentan auf Schüler scheißen | 57.69% | 50.00% | **38.46%** | 46.15% | 46.15% |
| Shareholder Value Investment Philosophie | **53.85%** | **53.85%** | **53.85%** | **53.85%** | **53.85%** |
| tagesschau 20:00 Uhr, 22.01.2020 | 20.00% | **16.00%** | **16.00%** | **16.00%** | **16.00%** |
| tagesschau 20:00 Uhr, 01.02.2020 | 6.25% | 6.25% | 6.25% | **3.13%** | 6.25% |
| tagesschau 20:00 Uhr, 03.02.2020 | **15.94%** | **15.94%** | **15.94%** | **15.94%** | **15.94%** |
| tagesschau 20:00 Uhr, 11.02.2020 | **11.76%** | **11.76%** | 17.65% | **11.76%** | **11.76%** |
| tagesthemen 22:15 Uhr, 18.02.2020 | **20.00%** | **20.00%** | **20.00%** | **20.00%** | **20.00%** |
| tagesschau 20:00 Uhr, 21.02.2020 | **12.50%** | **12.50%** | **12.50%** | **12.50%** | **12.50%** |
| tagesschau 20:00 Uhr, 03.03.2020 | **20.59%** | **20.59%** | **20.59%** | **20.59%** | **20.59%** |
| nachtmagazin 00:15 Uhr, 04.03.2020 | **17.74%** | **17.74%** | **17.74%** | **17.74%** | **17.74%** |
| Tagesschau in 100 Sekunden 05:17 Uhr. 04.03.2020 | **25.00%** | **25.00%** | **25.00%** | **25.00%** | **25.00%** |
| Venix - Corona Warn App | 26.00% | 26.00% | 26.00% | 25.00% | **24.00%** |
| Venix - TechNews 94 | **36.84%** | **36.84%** | **36.84%** | **36.84%** | **36.84%** |
| Venix - TechNews 95 | **45.31%** | 46.88% | **45.31%** | 46.88% | 48.44% |
| Venix - TechNews 96 | **47.14%** | 48.57% | **47.14%** | 48.57% | 50.00% |
| codecentric.AI Bootcamp - Was ist Machine Learning | 51.81% | 51.45% | 51.45% | **50.72%** | **50.72%** |
| **Total** | 39.50% | 38.40% | 38.47% | **38.33%** | 38.77% |

**Table A16:** EERs per file in test set “Anglicisms 2020” for all MTL models.