10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Preliminary guidelines for the efficient management of OOV words for spoken text

Christina Tånnander[1,2], *Jens Edlund*[2]

[1]Swedish Agency for Accessible Media
[2]KTH, Speech, Music and hearing
christina.tannander@mtm.se, edlund@speech.kth.se

Abstract

We investigate the practical short-term and long-term effects of five different frequency ranks used for selecting which out-of-vocabulary (OOV) words to add to a pronunciation lexicon for text-to-speech (TTS) of university textbooks. The work is an empirical study on a corpus of 200 university text books selected for talking book production and it takes the extensive pronunciation lexicon of a commercial text-to-speech system as its baseline. The main take-home message is a short but succinct set of guidelines that promise to increase the efficiency of OOV management, at least for text-to-speech production of university text books.

Index Terms: OOV, lexical coverage, pronunciation lexicon, speech synthesis

1. Introduction

Talking books are an increasingly important part of the broader accessibility effort as a result of the rapid performance improvements of text-to-speech (TTS) systems. More talking books than ever are produced with TTS, with considerably shorter production times. Today, the Swedish Agency for Accessible Media (MTM) provides towards a thousand talking university level text books yearly. With an average time from order to delivery of about 30 days (the actual TTS production is less than an hour), compared to 70 days average for human narrations, TTS offers invaluable possibilities from an accessibility perspective.

The quality of talking books produced with TTS improves steadily. These improvements are partially a result of the development of new and the honing of old TTS methods, but in real-world production, a key to quality is the management of out-of-vocabulary (OOV) words, the pronunciation of which continues to be a major stumbling block. Production systems still largely rely on a pronunciation lexicon – a lexicon mapping lexical words to a phonetic description of their realisation. Adding OOV words to the lexicon continues to be a straight-forward complement to automatic grapheme-to-phoneme (G2P) conversion that removes many quality issues and increases the overall listening experience. As the lexicon will never reach full coverage, it is also potentially infinitely time-consuming. Knowing how many words to add, when to add them, and how to select them is crucial.

The present work can be seen as a specific and goal-oriented application of corpus linguistics, which strives to provide a methodology for acquiring well-founded decision grounds in TTS development. We explore the effects of using different frequency-based ranks of OOV words, as a means of selecting which OOV words to add. The work presented is closely tied to the real production of talking books with TTS, and we maintain a strong focus on university text books. This is a quite specific type of written publication, but one that is highly important from a democratic and societal standpoint. The data – both the text corpus and the pronunciation lexicon – which we use to experiment, exemplify and illustrate, is all gathered from real-world production of such publications.

In a production setting, we can take two partially opposing views to quality assurance. With a short-term focus, we are interested in maximizing the immediate effects of a given effort. With a long-term focus, we raise our gaze to the horizon, disregard what is currently at hand, and look to maximize the overall effects over time of the same effort. Our goal here is to investigate both the long-term and short-term effects of different frequency-based selection strategies.

Finally, we acknowledge the caveat that not all OOV words are equally likely to be mispronounced, nor are all mispronunciations equally likely to cause listeners discomfort or comprehension problems. These likelihoods complement a comprehensive model for selection of which OOV words to prioritize. Although outside the scope of this paper, we note that such information can readily be merged with the corpus-based selection criteria we investigate here.

2. Background

2.1. Spoken text

This work is part of a larger project investigating various aspects of what has been referred to as spoken text [1], that is materials where text is the original form and speech is the realisation in which we take an interest. The distinction clarifies speech and language technology ambiguities and distinguishes spoken text from textual materials that are consumed as text (most language technology), speech materials that are consumed as speech (e.g. conversational AI, dialogue, any speech science that does not predominantly rely on transcripts), and speech materials that are consumed as text (transcriptions, "speech" science that mainly investigates transcripts). In spoken text, the written word is highly relevant, and text processing methods can be expected to apply.

2.2. Talking books

In discussions of read aloud books, one frequently encounters the legal distinction between talking books and audio books. Talking books are produced for people who cannot read printed text, often with an exception from the copyright law, while audio books normally refers to commercial books. The former is typically read aloud in a neutral manner, and the latter are often dramatized and produced with the intention of being entertaining. This distinction is far from new – in fact it was touched upon by Edison when he first described the

10.21437/SSW.2019-25

phonograph, the first machine that allowed us to record and replay speech. He states that books may be "read by the charitably-inclined professional reader" and be "used in the asylums of the blind, hospitals, the sick-chamber" (talking books), or for "great profit and amusement […] because of the greater enjoyment to be had from a book when read by an elocutionist than when read by the average reader" [2]. Here, our focus is on talking books.

2.3. The Swedish Agency for Accessible Media

The Swedish Agency for Accessible Media (MTM) is a government agency that produce literature in accessible formats such as Braille and talking books for people who for some reason cannot read printed text. The agency produces talking media in several genres: fiction, which is most often narrated by human voices, university text books, of which more than 50% are produced with synthetic speech (about 900 Swedish and English books/year), and more than 100 Swedish newspapers which are all produced with TTS [3]. Similar institutions exist in many countries, for example the British Royal National Institute of Blind People and the Library of Congress – National Library Service for the Blind and Physically Handicapped in the United States.

The sheer quantities of books produced makes it important to identify efficient production procedures. At the same time, the intelligibility of the talking media and the listener comfort they induce are crucial, in particular for text books which for some users are the only way to participate in education. This further emphasises the importance of good resource management, especially when it comes to cost intensive quality improvements.

2.4. University text books

The genres in which MTM produces talking books with TTS behave differently with respect to OOV words. In principle, the lexicon management for university text books can be done in advance, since the text is known well ahead of production. Conversely, news text affords much shorter lead times, and OOV words – which are often proper names and/or from another language than the target language [4] – are more likely be missing not only in the pronunciation lexicon, but in any available corpora as well [5]. Furthermore, talking books may be meaningfully resynthesized after lexicon improvements, whereas there is less demand for better readings of old news. The scope of this work is limited to talking university level text books, although the findings may to some extent apply to other genres as well.

2.4.1. TTS quality improvements

Here, we do not concern ourselves with quality improvements that are connected to changes in the TTS engine or its voices. Instead, we focus on quality assurance and improvements that can be implemented in the production line without any major changes to the TTS system, such as improvements to text preprocessing, lexicon improvements, and lexicon additions. Of these, the addition of new (OOV) items stands out as it is a task that is potentially endless. One way or another, any TTS production will have to make a selection of which OOV words to add. This work aims to make it easier to make an informed decision.

It should be noted that adding words to a pronunciation lexicon is not the only way to ensure that a certain word is pronounced correctly. In a language like Swedish, which is highly productive in compound formations, it can be enough to ensure that a specific word part is included in a set of possible compound parts. And automatic G2P conversion is a fall-back that at the very least ensures that each word gets some pronunciation. Although some G2P systems generate high quality pronunciations for some words, there is currently no system stable enough to replace the manual validation of unknown words entirely without an unacceptable loss of quality. Selection of which OOV words to add to the lexicon, then, is complementary to other improvements.

2.5. Efficiency matters

Adding OOV words in order of frequency to increase lexical coverage becomes more effortful the higher the lexical coverage is – this is an effect of the way words in corpora generally show a power-law distribution (e.g. [6], [7]). As the most high-frequent words are added, the remaining OOV words will be increasingly infrequent, and their addition will yield diminishing returns in terms of lexical coverage. As an example, the 1 000 most frequent lemmas of the Brown Corpus cover 72%, which is roughly 0.07% of coverage for each lemma added. For the next 5 000 lemmas, the absolute increase is 18% for a total of 90% coverage – a mere 0.004% of coverage for each lemma added [8]. In production-grade TTS, lexical coverage is significantly higher – roughly 95% in our data – and each addition requires manual inspection. Coping with this requires an efficient strategy for selection of which OOV words to add, and our main selection criteria is how effectively a strategy increases lexical coverage.

2.6. Previous work

2.6.1. Impact of OOV words

Much of the research on OOV words in speech technology is connected to speech-to-text, or automatic speech recognition (ASR). In TTS, effort has mainly been spent on methods to generate pronunciations for OOV words automatically.

2.6.2. Lexical coverage of specific word lists

There is a plethora of studies investigating the coverage of specific word lists on some given corpus. For example, the 2 000 most frequent word families from the General Service List (GSL) cover 75% of the tokens in non-fiction texts, and 90% in fiction [9].

2.6.3. Academic word lists

In the literature, we also find a range of studies concerning the selection of words to include in an *academic* word list, typically within the field of learning vocabularies in another language. The methods to achieve this range from using non-native English students' annotations of unknown words in English text books (e.g. [10]), to using reduced frequencies (e.g. [11]). Perhaps the most well-known English academic word list is the one that [12] compiled from a 35 million words corpus in 1998. The word list consists of 570 word families, selected by the following criteria: (1) the word family does not occur in the 2 000 first entries of *General Service List,* [13]*,* (2) the word family must occur at least ten times in each of the corpus' four main sub-sections, and (3) at least 100 times in the entire corpus. This word list covered about 10% of the corpus (and 8.5% of an unseen academic corpus), but only 1.4% of a fiction corpus of the same size, indicating that the selected word families actually are specific to academic text. Together with

the 2 000 word families from *GSL*, this academic word list covered 86% of the academic corpus. The Swedish Academic Word List [11] is built on a 2.5 million word academic corpus and consists of 750 lemmas covering 8.7% of the corpus.

2.6.4. Selection of words to add to a TTS lexicon

We are not aware of any structured investigations of the effects of different methods to select which OOV words to add for speech synthesis purposes. More generally, there is a number of methods to extend coverage of a lexicon. Methods range from the straightforward, such as adding English genitive forms of existing nominative nouns or proper names, or adding spelling variations of proper names [4], through using rhyme analogies to generate new pronunciations [14], to adding specific words that are extra likely to be mispronounced by a text-to-speech system, or to be misinterpreted by an ASR system.

2.6.5. Commonness and word frequencies

The most efficient way of adding words to a lexicon in order to increase its coverage of some corpus is to add them in order of frequency in that corpus. Although absolute frequencies are conceptually straightforward, they may not be the best sorting order to use if the target is to increase intelligibility. From a language learning point of view, the concept of **commonness** is sometimes used in lieu of absolute frequencies. Commonness is not necessarily simply the most high-frequent words in a corpus, and researchers, e.g. [11], [15], [16], have used adjusted metrics, such as reduced frequency, to include the distribution of the word in the corpus as a dimension to the commonness concept. A more accurate metrics for word types with low frequencies, average reduced frequency (ARF), is presented by [17]. Similar measurements have also been used by [15], for example average waiting time and average logarithmic distance. For a comprehensive overview of dispersions and adjusted frequencies, see [18].

Finally, the inverse document frequency (IDF), a measure of how specific a term is to a certain document in a corpus, was designed to help single out relevant index terms [19] by using it as a weight to frequency counts.

3. Method

3.1. Approach

We are looking to measure the effects of different word selection strategies in production systems. The contents of the pronunciation lexicon of such a system is typically compiled over long periods of time and under varying circumstances, and contains a lot of unpredictable, but not random, variation. Likewise, the materials to be synthesised are not randomly chosen, but guided by a number of factors such as current curriculums, student numbers, and individual student preferences. To capture this, we opt for an empirical approach in which we gauge effects on a representative, real-world data set.

The production of talking books is a continuous process, with well-defined sequential steps such as book selection, mark-up, pre-processing, text processing, TTS generation or reading by a human voice, post processing and distribution. In principle, quality assurance and improvements can take place in two manners: incrementally, as part of the production line of each book, or as part of system-wide upgrades, where a larger number of words are added each time. With incremental improvement, each addition takes immediate effect on the book currently in production, and over-all improvement is incremental. With system-wide updates, we can go back and re-synthesise books, but short of that, the improvements only effects books produced after the update. Our investigation captures these cases by gauging effects both on the current document (incremental updates) and on the corpus as a whole (both the results of a system-wide updates and as the end result of many incremental updates).

3.2. Data

3.2.1. Text corpus

The corpus MTMUNISV18 consists of 200 Swedish university books produced as talking books at MTM during 2018. The books come from different academic domains such as history, psychology, marketing, pedagogy and jurisprudence. The corpus is tokenised on word level, where each punctuation or white space delimited entity consisting of at least one letter is considered a word (i.e. digit-only sequences are not considered to be word tokens). The corpus consists of 15 166 894 such tokens, and the average token length is 5.95 characters.

3.2.2. Pronunciation lexicon

The baseline lexicon is a real-world pronunciation lexicon taken from the Swedish voice Ylva, created by Cereproc Ltd., and used in production by MTM. The lexicon contains 462 376 words. 352 773 of these are Swedish and 109 603 English. The high number of English entries is motivated by the frequent occurrence of English quotes, references, terms and other expressions in Swedish university text books. The lexicon also includes several proper names originating from other languages than Swedish or English.

3.3. Evaluation metrics

3.3.1. Relative lexical coverage improvement

Our main metric reflects how great an improvement to the lexical coverage a selection strategy will yield, given a specific number of additions. We calculate our **baseline lexical coverage** as the percent of our corpus that is covered by the baseline lexicon and use $\mathbf{COV_{CORP}}$ and $\mathbf{COV_{DOC}}$ for the improvements of lexical coverage after addition of words.

We use $\mathbf{COV_{CORP}}$ to denote the relative lexical coverage improvement of the corpus after addition of words. This is the quotient of the new lexical coverage and the baseline lexical coverage minus 1. Since the baseline coverage is high (95%), the relative and absolute lexical coverage improvements follow each other closely, thus only the former is reported here. $\mathbf{COV_{CORP}}$ represents long-term effects on the known corpus.

Similarly, $\mathbf{COV_{DOC}}$ denotes the average relative improvement over all single documents. Thus, the quotient of the coverage after the improvement under investigation and the quotient of the baseline lexical coverage minus 1 is calculated for each document. $\mathbf{COV_{DOC}}$ is the average of these 200 quotients.

3.4. Frequency ranks

We limit the investigation to frequency rank based selection, for the simple reason that with lexical coverage as the metric, frequency rank based methods will outperform other methods by nature. We investigate frequencies based on the entire corpus, on the one hand, and on the current document, on the other ($\mathbf{RANK_{CORP}}$ and $\mathbf{RANK_{DOC}}$, respectively). We compare three different selection strategies for frequency ranking:

absolute frequency, reduced frequency and term frequency-inverse document frequency.

3.4.1. Absolute frequency (RED)

Absolute frequency is frequency proper - the raw count of occurrences. We base the absolute rank either on the entire corpus (**RANK$_{ABS,CORP}$**) or on the current document (**RANK$_{ABS,DOC}$**). Adding the highest ranked words in **RANK$_{ABS,CORP}$** mathematically maximises **COV$_{CORP}$**, and adding the top ranked words in **RANK$_{ABS,DOC}$** maximises **COV$_{DOC}$.** In other words, these two methods make up the top-lines for corpus and document coverage, and the two additional ranks included here, reduced frequency and TFIDF, cannot beat them, but may have other advantages.

3.4.2. Reduced frequency (RED)

Absolute frequency does not distinguish between words that occur frequently in a small part of a corpus and those that are prevalent all over the corpus. Reduced frequency, in contrast, captures both frequency and dispersion. We follow [17] and calculate the reduced frequency ranks by finding the absolute frequency F for a word, dividing the corpus in F parts, and then counting the number of parts that contain the word. Analogous to the absolute frequencies, we base the ranks on the entire corpus on the one hand, **RANK$_{RED,CORP}$**, and on a single document on the other, **RANK$_{RED,DOC}$**.

3.4.3. Term frequency-inverse document frequency (TFIDF)

TFIDF reflects the importance of a word in a specific document, relative to the corpus as a whole. We calculate it according to [20], by calculating the absolute frequency F for a word in a document, and multiplying it by the inverse document frequency of the same word in the corpus. The inverse document frequency is the logarithmic function of the total number of documents divided by the number of documents that contain the term. TFIDF (**RANK$_{TFIDF}$**) is always calculated on the document level, with the entire corpus as reference.

3.5. Process

Relative lexical coverage improvements were calculated for the five different frequency rank-based selections shown in Table 1.

Table 1. *Rank-based selections for five types of frequency rank (rows) were evaluated for short-term (document) and long-term (corpus) effects on lexical coverage.*

	COV$_{DOC}$	**COV$_{CORP}$**
RANK$_{ABS,CORP}$	Average†	Batch††
RANK$_{RED,CORP}$	Average†	Batch††
RANK$_{ABS,DOC}$	Average†	Incremental†††
RANK$_{RED,DOC}$	Average†	Incremental†††
RANK$_{TFIDF}$	Average†	Incremental†††

3.5.1. Short-term effects

Short-term effects (**COV$_{DOC}$**) of *corpus- and document-based* frequency ranks were calculated by taking the average relative lexical coverage improvement for the addition of N = {1, 2, 4, 8, 16, 32} words for each of the 200 documents in the corpus (marked as † in Table 1). The metric estimates the immediate coverage improvement for a single book.

3.5.2. Long-term effects

Long-term effects (**COV$_{CORP}$**) of *corpus-based* frequency ranks were calculated by adding the top N*200 words from the corpus-based ranks of ABS and RED (TFIDF is not applicable at corpus level), and then calculating the improvement of the lexical coverage on the entire corpus (†† in the table).

Long-term effects (**COV$_{CORP}$**) of *document-based* frequency ranks were calculated by incrementally adding N words for each book in the corpus until reaching an addition of N*200 words, and then calculating the lexical coverage on the corpus (††† in the table).

These metrics estimate the corpus-wide coverage improvement, that is the average improvement for all future books.

For †††, we also calculated the incremental document coverage (**COV$_{DOC}$**) after each addition. The very first of these calculations is the equivalent of **COV$_{DOC}$** based on a single document, but as the process progresses, the coverage increases until the top ranked words of all 200 documents are added and we reach **COV$_{CORP}$**. The average of the numbers in-between, **COV$_{DOCPROC}$**, represents the average single document coverage throughout the process. The metric estimates how the coverage for the next books improves throughout the process.

Finally, we captured diminishing returns by dividing the coverage improvement for N={2,4,8,16,32) by N, to get the coverage increase per added word. We then normalise the results by dividing each result for N={2,4,8,16,32) with the result for N=1, to achieve a progression representing what percentage of the effect per word at N=1 we achieve when using Ns higher than 1.

4. Results

4.1. Type of frequency rank, short-term

Figure 1 shows the average relative increase in document lexical coverage (**COV$_{DOC}$**) after 32 words have been added using five different frequency ranks. The addition of 1, 2, 4, 8 and 16 words yields patterns of relative improvements that are near-identical to what is observed for the addition of 32 words in Figure 1, thus these results are omitted here for space reasons.

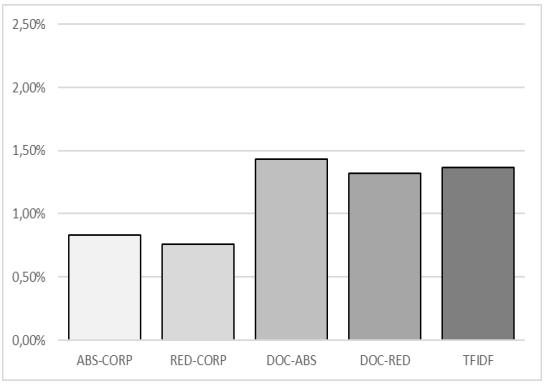

Figure 1. *Relative lexical coverage increase for a single document after addition of the 32 highest ranked words according to five frequency ranks (average over 200 documents).*

4.2. Type of frequency rank, long-term

Figure 2 shows the relative increase in corpus lexical coverage (**COV$_{CORP}$**) after 200 * 32 words have been added using five frequency ranks. Again, the relative performance of the frequency ranks is near-identical for the addition of 1, 2, 4, 8 and 16 words. Note that the coverage represented by the last three bars have been built up incrementally and that each document processed actually has a better coverage than the final corpus improvement shown in the figure. **COV$_{DOCPROC}$**, the average lexical coverage improvement for a single document after incremental OOV additions, show an improvement of 1.6%, 1.7%, and 1.6% respectively, for **RANK$_{TFID}$**, **RANK$_{ABS,DOC}$**, and **RANK$_{RED,DOC}$**.

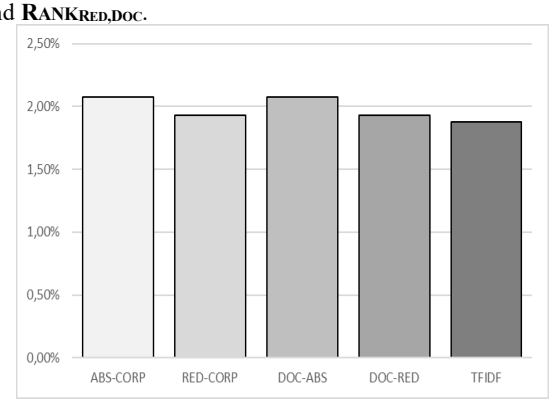

Figure 2. *Relative lexical coverage increase for the entire corpus after the addition of 200*32 highest ranked words according to five frequency ranks.*

4.3. Diminishing returns

The diminishing effect of adding a word can be modelled well by a power function for all methods. The methods can also be grouped into three groups based on their performance. For short term results, the first two and the last three methods in Figure 1 make two groups, and all long-term results can be grouped without significant loss of fit. Table 2 shows the statistics for these curves and Figure 3 shows the actual observations from first group together with the fitted curve.

Table 2. *Power functions for diminishing returns. Coverage increase at the N^{th} added word is N^A of the increase when N = 1. R square and residual for each curve is also included.*

Frequency ranks and coverage metric	A	R^2	RSS
RANK$_{ABS,CORP}$ **RANK$_{RED,CORP}$** Short term	-0.458	99.19%	0.0067
RANK$_{ABS,DOC}$ **RANK$_{RED,DOC}$** **RANK$_{TFIDF}$** Short term	-0.637	99.95%	0.0008
All ranks, Long term	-0.741	99.97%	0.0008

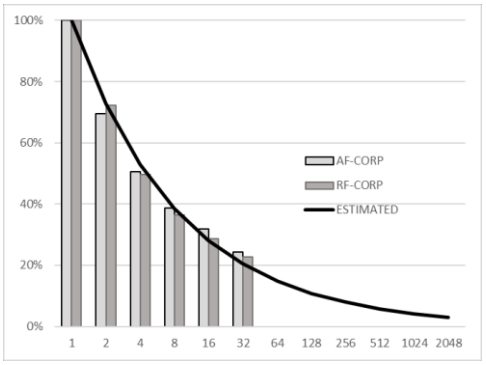

Figure 3. *Actual observations for two ranks (**RANK$_{ABS,CORP}$** and **RANK$_{RED,CORP}$**) with N={1,2,4,8,16,32} and the corresponding power curve.*

5. Discussion

5.1. Short-term considerations

The results of this study may seem bland, but on closer inspection, they provide pretty precise answers to the questions we set out with. We begin with the short-term perspective, where Figure 1 shows that using the corpus-based frequency ranks **RANK$_{ABS,CORP}$** and **RANK$_{RED,CORP}$** underperform at the single document level, compared to using document-based frequencies. Although we know that the document-based absolute frequency **RANK$_{ABS,CORP}$** performs best here by necessity, we note that the difference between corpus-based ranks and **RANK$_{ABS,CORP}$** is considerable, almost 2 to 1. On the other hand, we see in the same figure that the two document-based frequencies that take dispersion and term specificity into account (**RANK$_{RED,DOC}$** and **RANK$_{TFIDF}$**, respectively) both trail the top performer closely, and could be chosen as strategies at only a small penalty from a lexical coverage perspective.

5.2. Long-term considerations

Moving on to the long-term perspective, we note in Figure 2 that although the mathematical winner **RANK$_{ABS,CORP}$** indeed outperforms, the margins are very small all over. Its document-based counterpart **RANK$_{ABS,DOC}$** in particular does very well, and reaches an almost indistinguishable coverage. From this we suggest that when it comes to long-term effects, any of these frequency ranks will do from a lexical coverage perspective.

5.3. Consequences for the choice of process

We turn now to the question of what process is preferable for lexicon additions. We note that in the end, once 200*N words have been added, it does not matter much. But at the very beginning, when we add the first N words, it matters rather a lot, in that the corpus-based frequency ranks underperform severely. From a process standpoint, the corpus-based frequency ranks are easier to deal with in one go: we have the corpus and can add for example 6 400 words (32*200) in one big effort. This means, however, that *during* this effort (and it is a considerable effort), we get no reward. If, on the other hand, we build in the addition of N words as part of the preprocessing of each book and use a document-based frequency rank to select the words, we reach approximately the same long-term coverage once 200 books are produced, but the effect comes with the very first book. In fact, we get an average increase in document (i.e. book) lexical coverage of around 1.6-1.7%,

depending on which document-based frequency rank we chose. This is quite close to the final increase in corpus coverage after 200*N words are added, but we reap the rewards instantly.

5.4. More effort, less gain

The diminishing returns for all methods are readily modelled by power functions. The sharp drop, with exponents ranging from -0.74 to -0.46, is an expected result of distribution laws, but we see some variation that goes with method choice.

6. Conclusion

We feel that the current investigation, based on a corpus of 200 books and a production-grade pronunciation lexicon with more than 95% lexical coverage, constitutes solid grounds for a first set of guidelines for how to manage the addition of OOV words to a lexicon:

1. Make the addition of OOV words part of the production line and add a set number of words for each produced book, in the preprocessing stage. If the right selection strategy is used, this will make costing straightforward, and will have a direct effect on quality without losing out in the long-term.
2. Use a document-based frequency rank to select the words. If you believe that reduced frequency or TFIDF has important advantages over absolute frequency, then use them – the loss in lexical coverage increase is negligible.
3. Model diminishing returns for increasing numbers of added words and find a level that combines a reasonable cost with a decent result. The modelling is straightforward but depends on the frequency rank type to be used.

7. Future work

The present work focusses only on lexical coverage. There are other facets of OOV words that also influence how they affect quality – for example how difficult/irregular their pronunciation is, and how important it is that they be understood properly. It is our long-term goal to include such considerations in a decision model for OOV management, and the present work is a first step towards that.

We mean to clean up the code we used for this project and make it freely available. This will be a step towards gathering similar numbers based on different corpora and different pronunciation lexica. We will also provide support functions allowing questions like "how many words must I add to reach X% lexical coverage?" (by taking the answer from the diminishing returns estimations).

Finally, we again note that adding new words clearly is not the only lexicon management that will improve TTS quality. Pruning the lexicon might help reduce ambiguity and increase processing speed, and validating existing words and pronunciations is another task that would benefit from being managed efficiently. In the latter case, we believe that similar methods can be used to select which words to validate first in order to maximise effect, as well as to gauge the effects of pruning.

8. Acknowledgements

The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish research Council (2017-00626).

9. References

[1] C. Tånnander and J. Edlund, "First steps towards text profiling for speech synthesis," in *Proc. Digital Humanities in the Nordic Countries 2019 (DHN2019)*, 2019.

[2] T. A. Edison, "The phonograph and its future," *North Am. Rev.*, vol. 126, no. 262, pp. 527–536, 1878.

[3] C. Tånnander, "Speech Synthesis and evaluation at MTM," in *Proceedings of Fonetik*, 2018, pp. 75–80.

[4] J. Fackrell and W. Skut, "Improving pronunciation dictionary coverage of names by modelling spelling variation," in *Proceedings of the 5th Speech Synthesis Workshop*, 2004, pp. 121–126.

[5] M. Gerosa, M. Federico, and F.-I.-F. B. Kessler, "Coping with out-of-vocabulary words: Open versus huge vocabulary ASR," in *ICASSP*, 2009, pp. 4313–4316.

[6] F. Auerbach, "Das Gesetz der Bevölkerungskonzentration," *Petermanns Geogr. Mitt.*, vol. 59, pp. 74–76, 1913.

[7] G. K. Zipf, *Human behavior and the principle of least effort*. Reading, MA, USA: Addison-Wesley, 1949.

[8] B. Laufer and P. Nation, "Vocabulary Size and Use: Lexical Richness in L2 Written Production," *Appl. Linguist.*, vol. 16, no. 3, pp. 307–322, 1995.

[9] P. Nation and H. Kyongho, "Where would general service vocabulary stop and special purposes vocabulary begin?," *System*, vol. 23, no. 1, pp. 35–41, 1995.

[10] R. W. Lynn, "Preparing word-lists: a suggested method," *RELC J.*, vol. 4, no. 1, pp. 25–28, 1973.

[11] C. Carlund, S. Johansson Kokkinakis, J. Ribeck, H. Jansson, and J. Prentice, "An academic word list for Swedish-a support for language learners in higher education," in *SLTC*, 2012, pp. 20–27.

[12] A. Coxhead, "A new academic word List," *TESOL Q.*, vol. 34, no. 2, pp. 213–238, 2000.

[13] M. West, *A general service list of English words*. London: Longman, Green and Co., 1953.

[14] M. Y. Liberman and K. W. Church, *Text analysis and word pronunciation in text-to-speech synthesis*. New York: Marcel Dekker, 1992.

[15] P. Savický and J. Hlaváčová, "Measures of word commonness," *J. Quant. Linguist.*, vol. 9, no. 3, pp. 215–231, 2002.

[16] E. Volodina and S. J. Kokkinakis, "Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish," in *LREC*, 2012, pp. 1040–1046.

[17] J. Hlaváčová, "New approach to frequency dictionaries - Czech example," in *LREC*, 2006.

[18] S. Gries, "Dispersions and adjusted frequencies in corpora," *Int. J. Corpus Linguist.*, vol. 13, no. 4, pp. 403–437, 2008.

[19] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[20] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, 2004.