# First steps towards text profiling for speech synthesis

Christina Tånnander<sup>1,2</sup>and Jens Edlund<sup>2,3</sup>

<sup>1</sup> Swedish Agency for Accessible Media (MTM), Stockholm <sup>2</sup> KTH Speech, Music and Hearing, Stockholm <sup>3</sup> Språkbanken Tal, Stockholm christina.tannander@mtm.se, edlund@speech.kth.se

Abstract. We discuss an important yet under-studied domain of language and speech research: spoken text. Spoken text is language that was originally produced as text, then presented to recipients as speech. From a research perspective, this domain warrants special treatment, and we propose a classification that affords a structured approach based on a division of a linguistic message to be investigated into a primary (original) and secondary (studied) form. Secondly, we present the MTM Read Aloud corpus (MTM-RAC), a Swedish text and speech corpus built on in excess of 10,000 books. The corpus is closed access due to copyright restrictions on the material, but the methods developed and the results of our work on the corpus are available for use with similar corpora. MTM-RAC is designed with spoken text in mind and contains texts that have been read aloud in order to produce talking books, either by a human or using speech synthesis (i.e. text-to-speech) and the corresponding sound files. Finally, as the main purpose of the corpus is to explore and evaluate different aspects of text profiling for the purpose of reading aloud, we present first insights into this kind of profiling, based on experiments carried out on the corpus.

Keywords: read aloud text, spoken text, talking books, text profiling

# 1 Introduction

This paper inaugurates a long-term endeavour to untangle the relation between texts and the manner in which we read them aloud. Our chief motivation is to gain an understanding of how to better produce talking books with speech synthesis, or text-to-speech (TTS), but our methods and findings should be of interest to a wider audience.

We set out to achieve three goals here: Firstly, we provide a principled characterisation of *spoken text*, a domain of language that is both sufficiently different to other domains and internally coherent that it warrants separate treatment. Secondly, we describe a new Swedish corpus, the MTM Read Aloud Corpus (MTM-RAC). Although not freely available for copyright reasons, its existence will lead to common benefit in that it allows us to develop, validate and quantify new, freely available methods for profiling text for purposes of reading aloud. Thirdly, we present a first set of results as an example of one of the ways in which we will use MTM-RAC. The example is limited to one single measure, but one of the strongest candidate components for a more complete, complex profile: the relation between types and tokens as a text progresses.

On a more general note, examples of research questions we will be able to address by implementing and assessing text profiles of read aloud texts include: (1) What characteristics of the original text influence the characteristics of the corresponding read aloud speech? and (2) How can we make a computer read texts aloud in a humanlike manner, that is similar to how humans perform the same task? Follow-on questions include: (3) What speech characteristics influence how we understand the meaning of what is read aloud?; (4) How do different perspectives influence "quality" in read aloud text? For example, is read aloud text that is pleasant to listen to also intelligible, or are these orthogonal?; and (5) How are these questions influenced by text characteristics?

# 2 Background

This paper aims to (re)establish a field of research that is far from new, but exists in a vague space where a diverse set of disciplines blend speech and text into an opaque and nondescript haze. The first two sections of the background provided here is an attempt to untangle concepts and tease out a comprehensive description of our particular field of interest, and as such they are part of the research effort. We begin with a discussion of the relation between speech and text, and continue by throwing back to the origins of recorded speech, which lead up to the concept of talking books. The remainder of the background holds a discussion of text characteristics and brief overviews of the organizations behind the work and their motivations.

#### 2.1 Speech and text

With exception of fields that focus specifically on speech (e.g. speech technology and interactional phonetics), speech and text are often viewed as two sides of the same coin. When their differences are acknowledged, it tends to be superficially, and their treatment remain the same. For instance, language technology is purportedly an umbrella term for technologies dealing with text, speech, sign language, inter alia. In practice, the term is routinely used to denote quite precisely *text tools*. As an example, this year's Human Language Technology Conference, HLTCon2018, lists 14 different main topics, none of which has anything to do with speech, signs, of any other materials than text. At a glance, this may seem a harmless quirk. Superficially, speech and text are just two ways of encoding human language. In reality, the similarities between speech and text are not that clear, whereas the differences are striking.

The perhaps most fundamental difference is that speech is a natural language proper: it has evolved naturally in humans through use and without planning or premeditation. This does not hold for writing, which is a consciously designed language learnt under controlled forms in school. A common mistake is to take writing to be an encoding of speech. With the exception of pure speech transcripts, which are indeed an encoding of a subset of the information held in speech, this is not correct. The absolute majority of all writing is produced as text and intention to be read. Text is directed at one or more recipients, often unknown, that are not present at the creation of the text. It is intended for consumption at another time and place. It must, then, be self-contained, and it must be sufficiently clear that a reader will understand it without the affordance of questions. Speech, on the other hand, is produced in interaction with its recipients. A speaker can afford to be economical, and produce only as much as is needed, as there is continuous feedback from the recipients. Should the message get lost, clarification is available. Speech and text behave quite differently on just about every level, though. Text is wellstructured and grammatical, speech is dynamic and disorderly; text is unimodal, speech is multimodal; text is static and persistent, speech is emergent and transient. Consequentially, studies of speech proper requires very different methods than studies of text.

These considerations are often all but ignored. *Natural language processing* (NLP) routinely deals with written language – that is language that was originally produced in written form, not transcribed speech. Other fields, for example conversation analysis, deals with transcripts of speech, and others still, such as speech-in-interaction and speech technology include the speech signal itself as an object of study. Distinguishing clearly between the form of the source language and the form of the object of study allows a structured analysis (see Table 1).

**Table 1.** Language where text/speech is the original form (columns), held apart from the form in which they are studied (rows). Cell contents constitute examples and are not comprehensive.

		Primary form (original)	
		Text	Speech
ıry form died ation)	Written	Books, newspapers. Technical docu- mentation, journals, scientific writing.	Transcribed speech. ASR (automatic speech recognition) results.
Seconda (stu realis	Spoken	Read speech, scripted speech, TTS. Talking books and audiobooks.	Spoken interactions, unscripted presentations.

The categorisation is an idealized abstraction. There are clearly materials that fall between categories, and other examples are poorly matched in their category, such as real-time text chats, which would be categorised along with books. The categorisation does, however, cover many materials, and allows us to more readily describe the domains of several disciplines.

Written texts. Primary data for studies of properties of written language as they manifest in text. This is what NLP studies in practice and is the explicit primary data of a number of disciplines such as corpus linguistics and literature studies.

**Written speech.** Studies of the properties of spoken language as they manifest in transcriptions. NLP (in theory), conversation analysis, interaction analysis.

**Spoken speech.** Studies of the properties of spoken language as they manifest in speech. Speech in interaction, interactional phonetics, speech technology.

**Spoken text.** Studies of the properties of written language as they manifest in speech. Dramatics, theatre, performance.

In the "pure" cases, we study written texts as writing or speech as spoken data. Of the two mixed cases, studying speech as a written realization is the bread and butter of interaction analysis and conversation analysis, whereas studying text as spoken material is the domain of theatre and acting. The work we introduce here also fits in the spoken text category. Our aim, however, is to connect measurable characteristics of text with measurable characteristics of the reading of the same text, and we examine texts with a view to how their properties manifest in speech. We are less interested in the theatrical aspects, but focus instead on read aloud text as a means of delivering the text to those who cannot for various reasons read the printed words.

## 2.2 Talking books

140 years ago, when Thomas Edison patented the phonograph – presumably the first machine able not only to record sound, but to reproduce it as well – he so struggled with choosing the right name for the device that he gave it two names in his patent application: "phonograph" was one, the other "speaking machine" [1]. We can glean Edison's reason for dubbing a general audio recording device "speaking machine" from the North American Review article he published later the same year, in which he outlines his view on future applications of audio recording. In a list of eleven application areas, Edison lists dictation first. Next, before applications of music, toys, and memory aids, he proposes "Books":

"Books may be read by the charitably-inclined professional reader, or by such readers especially employed for that purpose, and the record of such book used in the asylums of the blind, hospitals, the sick-chamber, or even with great profit and amusement by the lady or gentleman whose eyes and hands may be otherwise employed; or, again, because of the greater enjoyment to be had from a book when read by an elocutionist than when read by the average reader." [2]

The passage impressively captures just about every aspect of talking books that you will see on the introductory slides of any present-day overview of the subject: the target audiences, the process of recording the books and who might do the work. To our knowledge, Edison's musings are the earliest clearly stated distinction between books read aloud for increased accessibility and books read aloud for increased enjoyment. The distinction has gained widespread use since, and many library services and authorities, 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, use *talking books* to denote the former and *audiobooks* to denote the latter. The difference between a talking book and an audiobook can be more than a technical peculiarity. In Sweden, there is also a legal aspect, as a talking book is produced with public funds and in accordance with Section 17 of the Swedish Copyright Act – a law that provides that permission from the holder of the copyright is not required to produce a published book as a talking book.

460

#### 2.3 Text profiles

There is a large body of work on readability measures of text (see e.g. [3]), mostly pertaining to how accessible a text is to a reader or a specific group of readers. Another, similar area looks at text quality from a more literary perspective. [4] lists a number of text quality indicators for Swedish prose (e.g. word, phrase, and sentence length; proportion of different types of punctuation marks). Looking at the way such metrics evolve as a text progresses is a way of creating a kind of profile of the text. In [5], the authors look at three parts of a single book from this perspective. The difficulty or ease with which a text can be turned into speech with speech synthesis is governed by a range of similar characteristics, from the proportion of new or unseen words and the proportion of foreign words and homographs, to the length and complexity of sentences, to the amount of tables and formulas in the text, to the difficulty level of the topic and the clarity of the writing. Our long-term aim, here, is to create simple, efficient and robust text profiles that allow us to predict the overall quality of a speech synthesis version of a text; to estimate the cost of reaching a certain quality; and to point to areas in the text where we would expect difficulties. The final version of these profiles will likely track deviations from expectations given by theoretical models of text (see [6] for a good overview). At this early stage, our chief interest is to examine the simple metrics that go into such models from an empirical point of view, to see what characteristics (e.g. text length, text genre) have a predictable effect on such models and thus should be controlled for.

**Types and type-token ratios.** The relation between types and tokens have been used for a wide range of purposes, and is the focus of several chapters in [6]. A simple means of investigating the relation is to calculate the ratio between the two. Youmans [7] makes a case against ratios, and argues for plotting raw values (e.g. type counts directly against token counts) in the following way: (a) type-token ratios in themselves, without relating them to the token count at which they are measured, are insignificant, and what is significant is instead "the rate at which they decline", and (b) plotting the ratio as a function of tokens is equally pointless, "since this ratio provides no more information than the raw data". Although the foundations laid forth by Youmans are true, there may be other compelling reasons to use the type-token ratio as a function of the token count. Youmans lists one of these reasons, but considers it a drawback: "the [type-token] ratio for any text (provided that it is sufficiently long) varies from a maximum of 1.0 to a theoretical minimum of zero". This, however, is a good property from a visualization point of view, as it is considerably more manageable to plot values that are known to vary between 0 and 1 than between 0 and infinity.

In the work presented here, we take a first look at the relation between types and tokens in MTM-RAC, with the goal of putting in place some guidelines that ensure that parameters that go into our profiles are not only sensible from a computational and modelling perspective, but also expressed in a manner that encourages visualization and examination by human analysts.

### 2.4 Swedish Agency for Accessible Media

The Swedish Agency for Accessible Media is a governmental authority that produces 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 areas: fiction, which is most often narrated by human voices, university text books, where more than 50% are produced with synthetic speech, as well as more than 100 newspapers produced with synthetic voices [8]. It is of great importance for the agency's users that the texts that are most suitable to be read by a synthetic voice are selected for the production with speech synthesis, while the least suitable books are recommended to be read by human narrators.

### 2.5 Nationella språkbanken and Språkbanken Tal

In 2017, the Swedish Science Council granted funding for a new national research infrastructure, *Nationella språkbanken*. The infrastructure is made up of three pillars: Språkbanken Text, with a focus on text-based language research, Språkbanken Sam, with a focus on societal aspects of language research, and Språkbanken Tal, with a focus on speech science and speech technology research. In addition, Nationella språkbanken became the administrator of Swe-Clarin, the Swedish membership in the European infrastructure Clarin ERIC. The speech infrastructure Språkbanken Tal was inaugurated in 2018, and is built from scratch. An early goal is to partake in resource and method development with external partners, as this will boost the build-up of resources. In that vein, Språkbanken Tal will make publicly and permanently available the methods and data that results from the work described here.

# 3 Method

### **3.1** The MTM Read Aloud Corpus (MTM-RAC)

We use a text corpus of 11,665 Swedish books in XML format that have been produced as talking books with text (as opposed to talking books consisting of speech only) at MTM. The material includes fiction and non-fiction directed at adults on the one hand, and young people and children on the other. The books are categorised into classes according to the Swedish library classification system, SAB [9]. Table 2 shows the number of books from each SAB class. Single letters represent literature for adults, and classes starting with *u* represent literature for children and young people. The class *uAV* does not exists in the classification system but is the result of merging all non-fiction books for children and young people. This was done because this subset is relatively small, consisting of 208 books all in all.

462

SAB	Description	#	SAB	Description	#
А	Books and libraries		Ν	Geography	93
В	General interest	147	0	Social science and jurispru- dence	1,628
С	Religion	177	Р	Technology, industry, and communication	118
D	Philosophy and psychology	410	Q	Economics and business	815
Е	Parenting and education	887	R	Sport, play, and games	99
F	Philology and linguistics	95	S	Military subjects	19
G	Literary science	109	Т	Mathematics	15
Н	Fiction	4,252	U	Natural science	100
Ι	Art, music, theatre, film, photography	260	V	Medicine	672
J	Archaeology	12	Х	Musical works, such as sheet music, piano rolls	9
K	K History		uH	Fiction for children and young adults	1,011
L	Biography with genealogy	213	uAV	Non-fiction for children and young adults	208
М	Ethnography, social anthro- pology, and ethnology	44			

Table 2. Proportions of the corpus based on SAB class.

#### 3.2 Process

**Text normalization and word tokenization.** Hyphens and quotes within words where normalized and delimiters such as punctuations and parentheses were deleted. All text was lowercased and tokenized in that sense a word was considered an entity between spaces or an opening or closing XML tag.

**Corpus subsets.** The corpus was divided into subsets of fiction and non-fiction for adults on the one hand, and for children and young people on the other (Table 3).

	Adults	Children and young people	SUM
Fiction	4,252	1,011	5,263
Non-fiction	6,194	208	6,402
SUM	10,446	1,209	11,665

Table 3. Proportions of the corpus on the highest level.

*Subsets by number of tokens I (SUBFIXBOOK).* The books for adults were further divided into equal-sized subsets based on the number of tokens in their body text. The fiction subset has five subsets of about 800 books, while the non-fiction subset has six subsets

of around 1,000 books. 33 fiction books with a token sum below 100 were excluded. No non-fiction book contained fewer tokens than 1,000 (see Table 4).

Total number of tokens in body text						
Subset	Fiction	Non-fiction	Subset	Fiction	Non-fiction	
1	100-5,500	1,000-29,700	4	44,000-85,800	57,000-73,700	
2	5,500-18,200	29,700-42,900	5	85,800-1,262,800	73,700-99,100	
3	18,200-44,000	42,900-57,000	6	-	99,100-881,500	

Table 4. Total number of tokens in the SUBFIXBOOK subsets of books for adults.

*Subsets by number of tokens II (SUBFIXLEN).* In addition, the books were split into fixed length subsets, based on their total number of tokens, resulting in the categories 1-5,000 tokens; 5-9,000;10-25,000; 25-50,000; 50-100,000; 100-200,000; and >200,000 tokens, with a varying number of books per subset.

*Subsets by SAB classification (SUBSABCLASS).* The books were also processed according to their SAB class, as explained in Table 2.

**Cumulative token counts, type counts, and type-token ratios.** Types and tokens were calculated cumulatively at every 100<sup>th</sup> token in every book. Results were truncated at 10,000 tokens, resulting in a list of 100 data points for types and tokens per book. Averages were then computed at each data point for each subset, and the type-token ratio was calculated at each point. We take as the type a graphic word, such that 'katt' and 'katter' ('cat' and 'cats') are different types. No consideration was taken to words that can be written in different ways, for example '24' and 'tjugofyra' ('twenty-four'). This is motivated by the aim to maintain simplicity and robustness and to avoid introducing new error sources. Only the body text was included in calculations in order to avoid undesirable consequences of for example tables of contents or registers, which could result in artificially high word type counts at the beginning or end of texts.

# 4 Results

**Raw counts vs. ratios and linear progression vs. logarithmic.** Fig. 1 presents a comparison of raw type counts as a function of token counts (the left column) and typetoken *ratios* as a function of token counts (right column) on the one hand, and a comparison of a linear representation of tokens on the X axis (top row) and a logarithmic representation of the same progression (bottom row). We note that (a) the linear typetoken ratio (upper right) clearly suggests that the ratio levels out as the book progresses, and provides a visual hint of the value at which this may take place, whereas the linear raw counts (upper left) are less easily interpreted, visually; (b) that the logarithmic representation of progression of token-type ratio through the book (lower right) expresses a near-linear relation (this can be verified: all curves over categories presented here



yield a good fit when described as logarithmic functions on the form TypeTokenRatio = A\*ln(TokenCount) + K); and (c) the number of types per token is consistently higher for the adult-directed texts.

**Fig. 1.** Four graphs of cumulative averages. Y is the average over all adult non-fiction books (solid line) and all child and young people non-fiction books (dashed line) at X tokens. The left column graphs show raw type counts on the Y axis, and right column show the type/token ratio. The upper graphs show raw token counts on the X axis; the lower show log counts on the X axis.

**Subsets by number of tokens (SUBFIXBOOK).** Fig. 2 shows that the total number of tokens in a book impacts its type-token ratio from the very beginnings of the text. Each curve represents one of the subsets in SUBFIXBOOK, and we note that after as little as 3,000 tokens, the subsets with higher total token counts show a higher type-token ratio (note the graph is zoomed in on both axes).

**Subsets by SAB classification, SubSABClass.** Fig. 3 shows the SAB classes with the lowest (parenting and education) and highest (geography) type-token ratios, together with the three intermediate SAB classes: medicine, biography with genealogy, and art, music, theatre, film, photography. Note that the Y axis starts at a 0.2.



**Fig. 2.** Truncated (3,000 < X < 10,000 tokens, Y > 0.23 types/token) type-token ratio curve for the five categories of adult fiction books (left pane) and six categories for adult non-fiction books (right pane) in SUBFIXBOOK. The lines arrange themselves in order of book size, with <5,500 tokens (solid line) at the bottom and >85,800 (dotted line) on top for fiction, and <29,700 (solid line) and >99,100 (dashed+dotted line) for non-fiction.



Fig. 3. Type-token ratio curves for selected SAB classes, Y starting at 0.2.

**Individual books.** Fig. 4, finally, shows type-token ratio curves for five single fiction books, evenly distributed by their total number of tokens in the body text (about 100, 200, 300, 400 and 500K). In the right-hand diagram, the X and Y axes have been truncated in the same way as in Fig. 2. to provide more detail. We note that the curves generally follow the same progression as the averaged curves in tables 1 through 4, but that the local variation is considerably higher.

466



**Fig. 4.** Five books evenly distributed according the total number of tokens: solid line ~100K tokens; long dashes ~200K; medium dashes ~300K; short dashes ~400K; and dotted line ~500K tokens. The diagram to the right is truncated for zooming effect.

# 5 Discussion

Youmans is correct in stating that type-token ratios fail to add information to raw type and token counts, but they improve visualization. Conversely, logarithmic representation of the token counts highlight the logarithmic progression of type-token ratio over tokens, but does not add to the visual clarity. Of the four visualizations of the data in Fig. 1, we prefer the linear progression over type-token ratios (upper right).

Looking at Fig. 2. we see that type-token ratios progression seems to depend on the total number of tokens in the book. This is a somewhat surprising finding. Although a higher total type count is expected in a longer book, it is less obvious that the difference is present in the beginnings of the text. We hypothesize that authors introduce many of the concepts in a text early on, leading to a higher initial type-token ratio when the final type count is higher. SUBFIXLEN, our second length based subset, shows the same pattern (graph not included here). We also note that the shortest fiction books category in Fig. 2. yield an uneven line. This is an artefact of the inclusion of books shorter than 5,500 tokens.

Fig. 3 (SUBSABCLASS) shows different progressions for SAB classes. N (geography) and I (art, music, theatre, film, photography) are the fastest, presumably due to the large variety of proper names, while E (parenting and education) and V (medicine) show lower ratios throughout. Medical literature typically holds a large proportion of unusual words (e.g. anatomical terms in Latin). Despite this, the type-token ratios are low, suggesting that words are similar among the 672 medical books in the corpus. L contains 213 biographies and genealogies with a large proportion of proper names, yet the type-token ratio is generally low. It may be that many of the 213 books within this class are biographies of a single person, rather than genealogies.

Plotting type-token ratio for single books (Fig. 4) predictably yields a more irregular progression. The irregularities in these lines illustrate what we believe is perhaps the

strongest read aloud indicator to be found in type-token ratio progressions: if we subtract the progression for the subset a book belongs to from the progression of the single book, we will acquire a horizontal line describing, at each point in the book, whether it currently has a type-token ration that is higher than, lower than, or similar to the average for the category.

## 6 Conclusions and next steps

In this work we have outlined a research area and placed it in a wider context. We have presented a text corpus that will allow us to make progress in the area, as well as some preliminary results pertaining to text profiling for text to be read aloud. The results immediately call for a more thorough investigation of the relationship between total book length and type-token ratio early in the text, as well as explorations of profiles that show the deviations of type-token progression in relation to class averages.

Another variation that we will look into at an early stage is the tokenization. Using simpler methods will make processing more efficient, and may not have a detrimental effect on results. We are looking to test truncation at 6 characters, a technique employed by many search engines, and tri-graphs (i.e. three-character sequences). We will also experiment with sample rates and with rolling windows of varying size.

Once we have a good hold on these basics, we will add more features such as parametric models to use as base-lines, and start looking for correlations between the profiles and speech characteristics in the corresponding read aloud texts.

## References

- [1] T. A. Edison, "Phonograph or Speaking Machine," 1878.
- [2] T. A. Edison, "The phonograph and its future," North Am. Rev., vol. 126, no. 262, pp. 527–536, 1878.
- [3] J. Falkenjack, K. H. Mühlenbock, and A. Jönsson, "Features indicating readability in Swedish text," in *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 2013, pp. 27–40.
- [4] L. Holm, "Rytm i romanprosa : en studie av rytmiska signalement i tio samtida svenska romaner," in *Det skönlitterära språket: tolv tetxter om stil*, C. Östman, Ed. Morfem, 2015, pp. 215–235.
- [5] C. Östman, S. Stymne, and J. Svedjedal, "Prose Rhythm in Narrative Fiction: the case of Karin Boye's Kallocain," in *Proc. Digital Humanities in the Nordic Countries 2018* (DHN2018), 2018.
- [6] R. H. Baayen, Word Frequency Distributions. Springer Science & Business Media BV, 2001.
- [7] G. Youmans, "Measuring Lexical Style and Competence: The TypeToken Vocabulary Curve," *Style*, vol. 24, no. 4, pp. 584–599, 1990.
- [8] C. Tånnander, "Speech Synthesis and evaluation at MTM," in *Proceedings of Fonetik*, 2018, pp. 75–80.
- [9] E. Viktorsson and M. Blomberg, "Klassifikationssystem för svenska bibliotek," 2015.