# Subtitles as a proxy for children's linguistic environment? Exploring the relationship between subtitles and speech in children's television programmes

*Maria Kjellholm*[1]*, Carla Wikse Barrow*[1]*, Sofia Strömbergsson*[1]
[1] *Division of Speech-Language Pathology, CLINTEC, Karolinska Institutet, Sweden*
maria.kjellholm@stud.ki.se, carla.barrow@stud.ki.se, sofia.strombergsson@ki.se

## Abstract

The study aimed to investigate the ways in which subtitles for Swedish children's programmes differ from the spoken dialogue they represent. Subtitles for 16 programmes, directed at two target groups (3-6 and 7-11 years), were compared to orthographic transcriptions of the corresponding spoken material, with regards to Mean Length of Utterance (MLU), Type Token Ratio (TTR) and lexical density. The results were further compared between the two target age groups. In accordance with expected patterns of children's language development the results show higher values for MLU and TTR for the older age group than for the younger. However, lexical density was, conversely and in opposition to expected findings, lower for the older children's programmes. The fact that these patterns were found not only in the spoken dialogue, but also in the subtitles, suggests that subtitles may be useful as a proxy for children's linguistic environment, for research regarding child language.

## Introduction

During the last ten years, there has been a growing interest in subtitles as a linguistic corpus. Such corpora already exist in many languages such as English, Polish and French (Mandera et al., 2015; New, et al., 2007; Van Heuven et al., 2014). Studies based on such subtitle-corpora have chiefly focused on the comparison of word frequency in corpora of subtitles and corpora of literary fiction (Van Heuven et al., 2014) or transcribed interviews (New et al., 2007), reporting significant correlations in word frequency between the two linguistic modalities. This lends support to the use of subtitle corpora in linguistic research of language at large.

Subtitles are, however, restricted in some respects that may limit their usefulness as a proxy for the spoken language they represent. One such restriction is the constraints regulated by the medium in which they occur; in order to fulfil their function, the subtitles must be presented simultaneously as the spoken dialogue. This restricts both their time and space on screen. Linguistically, these limitations involve reductions that take the form of omissions and/or substitutions (De Linde, 1995; Oksefjell-Ebeling, 2012). These processes can be assumed to reduce the dialogue to a more information-dense form. According to Oksefjell-Ebeling (2012), there is a strong correlation between the length of the spoken utterances and the amount of reductions in subtitles.

In their comparison between subtitles and literary fiction, Van Heuven and colleagues (2014) also included subtitles for children's TV programmes, revealing a significant correlation regarding word frequency between these subtitles and children's literature. Otherwise, subtitles for children's TV programmes remain relatively unexplored. Moreover, these researchers did not explore the correlation between subtitles and the spoken language they represent, thus mandating further research in this area. As language used by and directed to children generally can be assumed to be less complex than adult language, it is likely that subtitles for children's programmes should be subject to fewer omissions and substitutions than those for adult programmes. This suggests that subtitles for children's TV programmes may be more appropriate as a proxy than subtitles for adult-directed TV programmes for the spoken language they represent.

*Mean Length of Utterance*, MLU, is a standard measure which is often used in studies regarding children's language development (Hayes & Ahrens, 1988). MLU has been found to increase with age, both in the child's own language and in the language directed towards them, up to twelve years of age (Hayes & Ahrens, 1988). Another linguistic measure is *lexical density*, representing the information-density of an utterance (Johansson, 2009). Lexical density is estimated as the relation between words with lexical function and words with a primary grammatical function (Näsström, 2010). Lexical density has been found to increase with age, as children's linguistic abilities develop (Johansson, 2009). The traditional measurement for *lexical diversity* is Type Token Ratio (TTR). TTR is used to calculate the number of *different* words (types) in relation to the total number of words (token) in a given text. This measurement has also been found to be sensitive to children's language development, as language produced by and directed to older children (and adults) is associated with a higher TTR than language produced by and directed to younger children (Huttenlocher, 1998; Johansson, 2009). Notably, all three measures (MLU, lexical density and TTR) are also sensitive to differences in modality (Johansson, 2009); written language is distinguished by higher values in TTR and lexical density than spoken language.

Despite the fact the three measures (MLU, lexical diversity and lexical density) illuminate different linguistic aspects, they are often found to correlate. For example, lexical density and lexical diversity have been proposed as useful indicators of lexical development and correlate strongly with one another (Johansson, 2009). As mentioned above, this lexical development is paralleled by an increase in MLU (Hayes & Ahrens, 1988). Regarding the measures' sensitivity to differences between subtitles and the spoken language they represent, less is known. Due to the undefined nature of the linguistic form of subtitles, and as it cannot be assumed to be purely of either the spoken or written form of language, predictions of the measures' values cannot be based on studies of either of these language forms. However, due to the reductions previously mentioned (Oksefjell-Ebeling, 2012), lexical density and diversity can be assumed to be higher and MLU to be lower in subtitles, compared to spoken language.

The aim of this investigation was to examine how and to what degree subtitles of Swedish children's TV programmes differ from the spoken dialogue they represent. Specifically, the research question was whether subtitles and spoken dialogue differ in regards to utterance length (MLU), lexical density and lexical diversity (TTR) and also, if any difference is dependent on the target group of the programmes. Based on prior research on subtitles (Oksefjell-Ebeling, 2012) and children's developing linguistic environment (Hayes & Ahrens, 1988; Huttenlocher, 1998) the hypothesis was that subtitles and the spoken dialogue differ with regards to MLU, lexical density and lexical diversity in such a way that subtitles contain shorter utterances, are more lexically diverse and more lexically dense. All linguistic measurements were hypothesised to be higher, and the difference between modalities to consistently be larger for the older age group.

The knowledge gained from this investigation regarding the relationship between subtitles and corresponding speech will address a current void in this research area. Good agreement between subtitles and spoken dialogue could merit the use of subtitles as a proxy for children's linguistic environment, where available resources are still scarce. The linguistic domain of (public service) television is attractive because it reaches households across nations, thus exposing hundreds of thousands of children to the same linguistic input.

## Method

Subtitles for 56 episodes of 16 children's programmes of mixed genre, and orthographic transcriptions of the same episodes were included in this study. Of the 16 programs, 8 were from the target group 3-6 years and 8 from the target group 7-11 years, as defined by SVT. In total 32 episodes from the younger age group and 24 episodes from the older age group were used. Calculations of linguistic measurements were automatically performed with a Perl script, and SPSS was used for statistical analysis of the data.

The programmes were selected based on popularity and availability on svtplay.se. From each programme, 15 minutes were orthographically transcribed by four SLP students. Depending on the length of the episode, the minutes were evenly distributed across three or five episodes. The subtitles for the selected programmes, all of which had been produced manually, were provided by SVT Undertexter, who bear the formal responsibility for all subtitles produced for SVT.

The samples were orthographically transcribed, whilst playing the episode on svtplay.se. All through the transcriptions, non-human speech (e.g. ”vov vov”) and non-linguistic sounds (e.g. laughter) were omitted. From the transcribed material, a sample of 30% was re-transcribed by another transcriber to allow estimation of consistency. After a review of the re-transcriptions, certain revisions were made in the original transcriptions, to ensure a consistent use of the transcription guidelines. To increase the congruence with the subtitles, without compromising the integrity of the transcriptions, some well-established reductions were transcribed with colloquial spelling (e.g. "något" as "nåt"). Linguistic elements of the subtitles not realized in the spoken material (e.g. "SKRIK" or "ÅSKA") were omitted before analysis.

The analysis of the correlation between the original transcriptions and the re-transcriptions, as measured by a Pearson's correlation analysis, revealed very strong correlations with regards to MLU (r = .939, n = 54, p <.001), to lexical density (r = .974, n = 54, p <. 001), and to lexical diversity (r = .993, n = 54, p <.001).

The calculation of lexical density was based on a list of function words from the 100 most frequent words in the entire material. Following the recommendation by Watkins and colleagues (1995), lexical diversity (TTR) was calculated from equally sized samples consisting of the 100 first words from each transcription. A two-way ANOVA was conducted, with the three linguistic measures as dependent variables and age group and modality (i.e. subtitle vs. transcription) as independent variables.

## Results

As seen in *Table 1*, the examined 240 minutes of television resulted in more utterances as represented as speech, than as represented in the subtitles. The table further shows that, although of the same time length, programmes directed towards older children contained more utterances.

*Table 1.* Number of utterances, as distributed across target age groups and modalities.

|  | Number of utterances |  |
|---|---|---|
| Target group | In subtitles | In speech |
| 3-6 ys | 1714 | 1896 |
| 7-11 ys | 1827 | 1974 |

As indicated in Figure 1, MLU was consistently shorter in spoken dialogue than in subtitles, F(1,108) = 48.69, p <. 001, $\eta^2$ = .31. The figure also illustrates the significant difference found between age groups, F(1,108) = 20.02, p <. 001, $\eta^2$ = .16. A significant interaction effect was found between age group and modality, F(1,108) = 6.32, p <. 013, $\eta^2$ = .06. This effect is also illustrated in Figure 1, as an increased difference in MLU between modalities for the older age group compared to the younger.

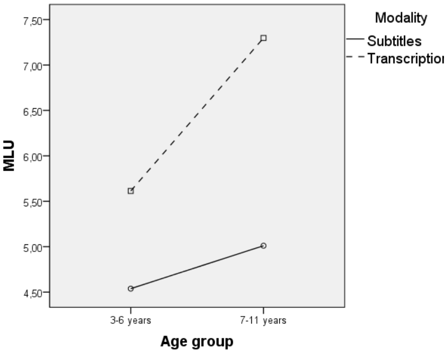

*Figure 1.* MLU for target age group and modality.

As observed in Figure 2, lexical density was higher within the younger target group, than within the older target group; F(1,108) = 7.71, p <. 006, $\eta^2$ = .07. However, there was no significant effect of modality: F(1,108) = 3.23, p = .075, nor was the tendency to interaction between age and modality significant: F(1,108) = .25, p = .620.

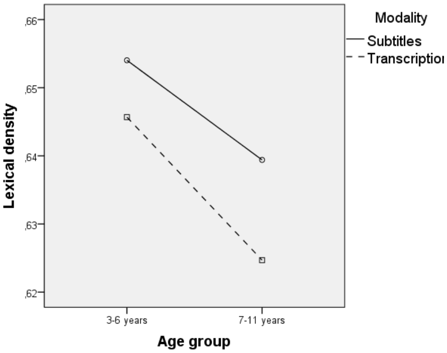

*Figure 2.* Lexical density for target age group and modality.

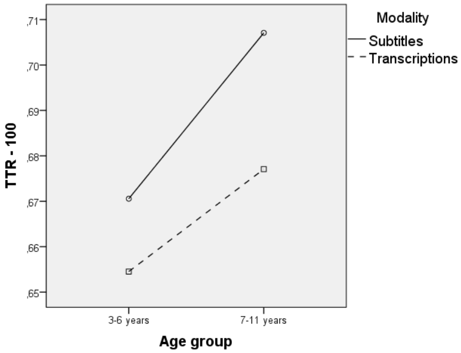

*Figure 3.* Lexical diversity for target age group and modality.

Figure 3 indicates that lexical diversity (TTR) was higher in subtitles than in speech. This tendency was not, however, significant: F(1,108) = 3.25, p=.074. Regarding age group, on the other hand, a significant main effect was found: F(1,108) = 5.35, p = .023, $\eta^2$ = .05, such that lexical diversity was higher in programmes directed towards the older age group. The tendency to interaction between age and modality was not significant: F(1,108) = .299, p = .586.

## Discussion

The results presented in this study show that MLU was significantly lower in subtitles compared to the speech they represent. Non-significant tendencies were found for higher lexical density and TTR in the subtitle modality, thereby only partly confirming the expected difference between the two modalities. Furthermore, the results revealed that MLU and TTR were lower and lexical density higher, in both subtitles and spoken dialogue, in programmes directed towards the younger target group compared to programmes directed to the older age group.

Across all three linguistic measures, a tendency regarding differences between subtitles and spoken material was observed as being larger within the older target group than within the younger (although admittedly, this interaction was significant only in the case of MLU). This pattern suggests that subtitles for programmes directed at younger children could function better as proxy for their linguistic environment than those for programmes with an older target group.

In written language, higher values of TTR and lexical density are expected compared to spoken language, based on previous research (Johansson, 2009). Although a tendency in this direction was observed in the present investigation, the difference between subtitles and spoken language with regards to these measures was not significant. This may indicate that lexically, subtitles bear more resemblance to spoken language than to written language in general.

The observed pattern of increased MLU and TTR with growing age resonates with previous research based on spoken and written language in general (Huttenlocher, 1998; Johansson, 2009). The fact that the same pattern is observed also for subtitles indicates that they reflect similar linguistic adjustments to age as previously studied forms of language. This may serve as support for the suggested use of subtitles as a proxy for children's linguistic environment. Notably, however, the relation between language directed towards the two different age groups in regards to lexical density diverts from previous research, as this measure is expected to increase with age and linguistic ability (Johansson, 2009). A possible explanation to why lexical density is higher for the younger age group than for the older may be sought in the language usage of the programmes directed towards the respective age groups. Closer inspection of the programmes produced for the older target group revealed that these, to a larger extent than those directed at younger children, included natural speech and dialogue, with interrupted and/or simultaneous speaker turns as well as a high density of function words. The programmes produced for the younger group, on the other hand, were often scripted, with an educational, varied and topic-specific vocabulary that included very few interruptions. It may be that given a larger data-set, with a broader coverage of episodes and programmes, the analysis of lexical density would yield more anticipated results.

The size and coverage of the material included for analysis in the present study was restricted by a rather limited time frame. Moreover, the selection of programmes as well as the definition of age groups was restricted by what was available and defined by SVT. This excluded the possibility of the authors having to make subjective selections of programmes. However, it also limits the granularity of the analysis with regards to age groups, as only two broad target age groups were defined. It may be desirable in future research to define target age groups more narrowly, in order to allow more fine-grained analysis of the potential effects of age on the linguistic nature of the subtitles. Moreover, and as mentioned above, future research should strive for a broader coverage of children's programmes than what could be included in the present study.

As programmes directed towards younger children are often shorter than those directed towards older children, certain limitations arose concerning transcription time per episode. Accordingly, it was necessary to transcribe three minutes of five episodes rather than five minutes of three episodes for certain programmes in the younger target group. Whether this affected the reliability of the measurements could not be determined.

Given the promising results presented in this study, the method and results can be used to inspire more comprehensive research into subtitles and the spoken dialogue they represent. More specific inspection into what linguistic elements that are omitted, reduced and substituted as well as to what degree, are necessary for a fuller understanding of the linguistic nature of subtitles. Whether the results presented in this study give evidence for the use of subtitles as a proxy for children's linguistic input should be further investigated, ideally by comparing a larger pool of subtitles to existing corpora of natural language input directed towards children.

There are many linguistic aspects that remain unexplored using the presented methodology, e.g. morphological and syntactic structure. The measurements used in the study do, however, also reflect certain parts of these aspects and can be regarded as a rough estimation of linguistic complexity. The fact that the linguistic measures consistently show different values for the two age groups is a clear indication that they are sensitive to linguistic variation in language directed towards children of differing age. That this difference is observed not only in the spoken dialogue but also in the written representation of the dialogue, suggests that subtitles could function as a substantial resource in future child language research.

## Acknowledgements

This study was based upon a course assignment, at the Speech and Language Pathology program at KI, conducted by the first two authors in collaboration with Ellinor Finne and Isabell Olsson; "Undertexter som korpus? - en jämförelse av text och tal i svenska barnprogram". The original report was translated and reworked by the authors of this paper.

## References

De Linde, Z. (1995). 'Read my lips': subtitling principles, practices, and problems. *Perspectives: Studies in translatology, 3*(1), 9-20.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'?. *Journal of child language*, *15*(02), 395-410.

Huttenlocher, J. (1998). Language input and language growth. *Preventive Medicine*, *27*(2), 195-199.

Johansson, V. (2009). *Developmental aspects of text production in writing and speech.* Lund: Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University, 2009.

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods, 47*(2), 471-83.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, *28*(04), 661-677.

Näsström, S. (2010). *Lexikala och syntaktiska aspekter i skriftligt berättande hos barn med språkstörning* (Magisteruppsats). Lund: Institutionen för kliniska vetenskaper, Lunds Universitet. Available at: https://lup.lub.lu.se/student-papers/search/publication/2862379

Van Heuven, W., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity differentiating Typical and Impaired Language Learners. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1349-1355.