Received: 23.03.2020
Accepted: 27.07.2020

DOI: 10.23817/strans.11-13

Andy Stauder, Michael Ustaszewski
University of Innsbruck/ Austria

Syntactic complexity as a stylistic feature of subtitles

Abstract

Syntactic complexity as a stylistic feature of subtitles

In audiovisual translation, stylometry can be used to measure formal-aesthetic fidelity. We present a corpus-based measure of syntactic complexity as a feature of language style. The methodology considers hierarchical dimensions of syntactic complexity, using syllable counting and dependency parsing. The test material are dialogues of several characters from the TV show *Two and a Half Men*. The results show that characters do not differ syntactically among themselves as much as might be expected, and that, despite a general tendency to level differences even more in translation, the changes in syntactic complexity between the original and translation depend mostly on the respective character-feature combination.

Keywords: translation style, stylometry, subtitles, syntactic complexity, corpus approaches

1. Introduction

Linguistic style has been a popular topic for a while, although a widely adopted definition of the concept is lacking. There seems to be a notion of *choice of linguistic expression*. The concept is similar to that of *translation quality* – quality literally meaning "how-ness", which is similarly poorly defined and thus has elicited what House (1977: 1) calls "anecdotal, biographical and neohermeneutic approaches to judging translation quality", i.e. approaches that have hardly anything to say about linguistic features in the narrow sense of the word.

A more objective and much less vague approach is the seminal corpus-based methodology by Baker (2000). This proposes a clearly defined set of measurable features for describing what is called *translator style*: frequencies of certain words or parts of speech, mean sentence length, standardized type-token ratio,

etc., "typical of a [given] translator" (ib.: 245). From this, it is clear that style can be layered: the style of the original may be changed in a consistent fashion by the translator. Also, style does not necessarily have to consist of conscious choices, but may be habitual and due to a variety of influencing factors. These can be described very well with the *diasystem* by Coşeriu (cf. 1981[1958], Goossens 1977, Faust 1988), which classifies linguistic modes of expression according to social context. Thus, a person's mode of linguistic expression may be influenced by class, age, geographical provenance, historical period, and target audience/situation. All of these may also be fictitious: an author may *want* to have one of their characters sound a certain way. This adds a third layer of style: on top of the personal one of the author, which may be influenced by the aforementioned *dia* factors, comes the one of the possible characters created by the author, which may then be superseded by the alterations due to the translator's style, which may again be influenced by the *dia* factors and conscious choices of the translator.

So, on the one hand, style characterizes the way a person writes and translates, and, on the other, can also be used as a creative device for writers (or translators) to shape the characters in their works. This is especially true for audiovisual entertainment: this is usually very character-heavy and one main feature characterizing the protagonists is the way they talk. The goal of this research is therefore the following: it is interested in operationalizing one feature of linguistic style – syntactic complexity – for Audiovisual Translation, with the results having possible application in the identification of translator style and also translation quality, at least as far as similarity of source and target text are concerned. Thus, the paper's aim could be said to fall within the area of *translation stylometry* (cf. Lynch 2017; Rybicki 2006; Rybicki 2012). The research is to be conducted on subtitles because these present audiovisual language data in a form that lends itself to machine-processing.

There are few studies that specifically target syntactic complexity as a stylistic feature: most seem to focus on lexical, i.e. semantic and pragmatic (cf. e.g. Kenny 2001; Winters 2007; Saldanha 2011), but not so much syntactic phenomena. Those that do seem to do so indirectly, by examining the feature of readability, and by applying it to whole texts rather than individual sentences/utterances (cf. e.g. the insightful Huang 2015: 95 ff.). According to Huang (2015: 115) "It is found that statistics about readability provided by manual calculation or computer software cannot effectively differentiate one text from another in terms of style." This study, on the other hand, looks to pin down syntactic complexity as a stylistic feature of individual utterances, in the context of the dialogue-heavy field of Audiovisual Translation. Here, as in any (quasi-)literary work, it is not only important what is being said, but also the way it is being said (cf. what Jakobson 1972[1960] called the *secondary structure* of literary texts). Therefore, the research questions are:

A) Is syntactic complexity a meaningful stylistic feature of linguistic utterances in the form of subtitles, i.e., do the characters of a TV show discernibly differ from each other with regard to this feature?
B) To what extent is the syntactic complexity of a specific TV show's characters' utterances reproduced in their translation, i.e., by how much do complexity scores change?

While finding answers to these research questions, the study also aims to tackle a problem that may not be relevant to readability studies, but which is to translation studies: the scores of linguistic complexity calculated for the subtitles from the test corpus are to be adjusted for language-specific variation with the help of representative corpora. The reason for this is that a specific English sentence may, in comparison to the English language in general, be more complex than its German translation in comparison to the German language in general.

The methodology for measuring syntactic complexity is one devised by the authors and takes into account word count per sentence, syllable count per word, and dependency tree complexity. The reason for devising a dedicated methodology is that classical readability or syntactic complexity calculation methods are somewhat limited and may not capture the phenomenon of syntactic complexity adequately: they mostly limit themselves to linear features (word and sentence length; cf. e.g. Flesch 1948; Björnsson 1968; Björnsson 1983) or account for structural information only heuristically, e.g. by including verb count (cf. Fichtner 1981).

The test material consists of a parallel corpus compiled by the authors, containing the English and German subtitles of the first season of the show *Two and a Half Men*, with a size of approximately 60,000 words in total. The corpora for standardizing the scores are the German and English versions of W2C (Majliš/Žabokrtský 2012).

2. Research Design

The first thing that was required for attempting to find answers to the aforementioned research questions was the gathering and preparing of data. The material in question was the first season of the TV show *Two and a Half Men* (Warner 2012). The subtitles were sourced from the German-language 8-season box set of the show (ib.) using the built-in OCR function of the free software *SubtitleEdit* (Olsson 2019). After correcting a considerable number of OCR errors, the subtitles were character-tagged manually with the help of the same software, using tags consisting of a pound sign and two letters each, e.g. #CH for Charlie Harper, one of the show's main protagonists, portrayed by Charlie Sheen.

The next step consisted in performing basic parsing of the material. For this, the *udpipe* package (Wijffels 2018) for the R programming language was used. This was used for tokenizing the material and parsing the dependency trees of

all the contained sentences. Based on this, the features used for assessing the syntactic complexity of each sentence (and ultimately the parlance of each character of the show) could be calculated. The features used (and calculated for each sentence) were the following:

- word count,
- average syllable count per word,
- hierarchical complexity: edges per leaf (in the dependency-parsed tree).

The word count could be calculated directly using R, as the data had been tokenized before. The syllable count of each word, which was the basis for determining the average syllable count per word of each sentence (also using R), was calculated using two R packages: *quanteda* (Benoit et al. 2017) for the English subtitles, and *hyphenatr* (Rudis 2016) for the German subtitles because quanteda had been found to perform poorly for German. This is probably due to the higher degree of inflection of the German language, which leads to a larger number of types per a given number of tokens and consequently the need to develop a more refined approach as compared to a mostly isolating language such as English. In terms of a hierarchical dimension of syntactic complexity, which is missing from most approaches to measuring syntactic complexity as discussed above, a rather simple but effective approach has been chosen. The simple part of the approach is, in fact, the last step of the calculation. This consists in counting how many edges per number of leaves there are in the respective syntactic (dependency) tree of a given sentence. Figure 1 shows this concept: two trees may have the same number of leaves (i.e., terminal nodes, represented by hollow circles in the figure), but may differ as far as their numbers of edges (connecting the nodes) are concerned. This method provides a computationally efficient way of assessing what the typical path depth within a tree is.

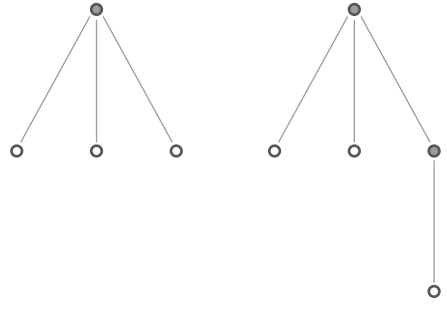

Figure 1. Trees with the same number of leaves (here in white), but more nodes, are more complex.

The first step consisted in calculating the values for each of these three features for each of the sentences, in both languages: German and English. Then, the mean

value for each character from the show was calculated for each of the three features by summing up all the calculated values of all the sentences of a character and then dividing by the number of sentences of that character. The typical values for each language might reasonably vary: e.g., the typical sentence length of a German sentence can be assumed to be greater than that of the typical English sentence and judging a German sentence of a length of say, ten words, in the same way as an English one of the same length would be unfair. Therefore, to bring the character-specific values to the same scale, the same calculations as for the test corpus, i.e., the German and English subtitle sentences, were made for each sentence in a reference corpus for each language and, again, their mean was calculated. The reference corpora were the German and English version of W2C (Majliš/ Žabokrtský 2012). For the evaluation of the test corpora, the calculated feature values for each character of the show where expressed as fractions of the form

test corpus value ÷ reference corpus value

such that, e.g., a character speaking, on average, using sentences of 40% of the length of the average length of sentences in the reference corpus for the respective language would be assigned the value of 0.4. This calculation was performed for each of the seven main characters of the show, for each of the three aforementioned categories: word count per sentence, dependency complexity expressed as the number of edges per leaf of a sentence's dependency tree, and average syllable count of all the words in a sentence.

3. Results

Table 1 shows the results for each main character of the show for both languages. The characters are: AH – Alan Harper; BE – Berta; CH – Charlie Harper; EH – Edith Harper; JH – Jake Harper; JU – Judith (Harper, then Melnick); RO – Rose.

Table 1: The calculated values for each character and both languages (rounded to two decimal places)

Character	wc_en	wc_de		dc_en	dc_de		sc_en	sc_de
AH	0.38	0.39		0.85	0.87		0.82	0.90
BE	0.42	0.42		0.88	0.86		0.82	0.92
CH	0.38	0.38		0.84	0.87		0.81	0.89
EH	0.43	0.46		0.87	0.88		0.85	0.87
JH	0.31	0.31		0.83	0.89		0.79	0.89
JU	0.42	0.43		0.87	0.90		0.85	0.89
RO	0.39	0.41		0.85	0.89		0.84	0.89

The values are (_en for English, _de for German): wc – average word count of a character's sentences; dc – average dependency complexity (=edges per leaf) of a character's sentences; sc – mean of all average syllable counts per word of each of a character's sentences.

Figure 2 shows the changes in complexity that happened during the translation in graphical form.

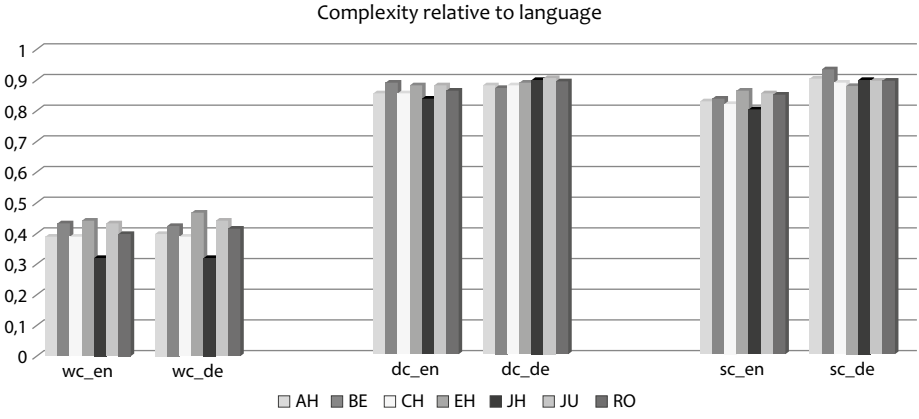

Figure 2: Changes in complexity during translation

Based on these data, Research Question A): "Is syntactic complexity a meaningful stylistic feature of linguistic utterances in the form of subtitles, i.e., do the characters of a TV discernibly differ from each other with regard to this feature?" can be answered, by looking at the degree to which the values of each character differ from those of the other characters. A tool that helps assess this variation between characters more summarily than a mere comparison of the individual values is the calculation of the coefficient of variation for each feature and each language. These coefficients of variation (rounded to the first decimal) are shown in Table 2.

Table 2: Coefficients of variation of syntactic complexity features

Language	wc	dc	sc
en	11.0% (σ: 0.043, μ: 0.391)	2.1% (σ: 0.019, μ: 0.856)	2.5% (σ: 0.021, μ: 0.828)
de	11.9% (σ: 0.047, μ: 0.398)	1.4% (σ: 0.013, μ: 0.881)	1.8% (σ: 0.016, μ: 0.892)

Research Question A can thus be answered as follows: there is noticeable variation between the various characters of the studied show when it comes to the

syntactic complexity of their utterances, but its degree is dependent on the type of feature in question. The word count per sentence varies the most, with a coefficient of variation of more than 10% in both languages; the average syllable count per word of each sentence varies considerably less, at a variation on the order of 2% with a more pronounced degree in the English original than in the German translation. The least variation could be found with regard to dependency complexity (edges per leaf of the dependency tree of a sentence), with values of barely 2% for the English original and well below 2% for the German translation; so, again a reduced degree of variation in the translated text.

The second indicator for which data was gathered and processed is the difference in the syntactic fidelity of the translations, i.e., in how far the syntactic complexity of the original, quantified by the three factors discussed, changed in the respective translation. The average changes from the English original to the German translations are listed, in the form of percentages, in Table 3.

Table 3: syntactic fidelity – percentages of change in syntactic complexity between English original and German translation

Character	wc (%)	dc (%)	sc (%)
AH	1.68	3.17	9.06
BE	-1.28	-1.94	12.05
CH	0.43	3.37	8.99
EH	6.21	1.07	2.12
JH	0.51	7.65	1.12
JU	0.84	3.21	1.05
RO	4.22	3.85	5.55
σ	2.56	2.90	3.78
μ	1.80	2.91	7.89

The value categories (wc, dc, and sc) and character designations (AH, BE, etc.) are the same as the ones described for Table 1. In the following, Figure 3 shows these differences in the form of a bar chart. The scale is decimal, not in percentages.

These data help answer Research Question B): "In how far is the syntactic complexity of a specific TV show's characters' utterances reproduced in their translation, i.e., by how much do complexity scores change?".

As can be seen from the presented data, the change varies depending on the respective feature-character combination, ranging from -1.28% to 12.05%. The general tendency for this show and this language pair is an upwards one:

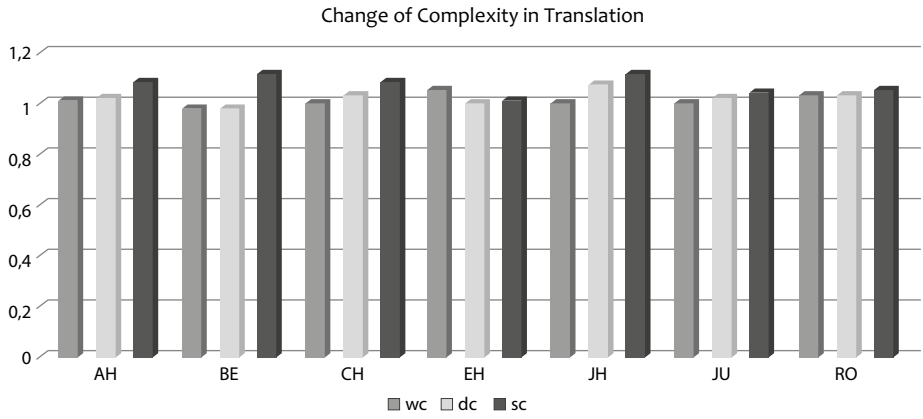

Figure 3: Complexity changes in translation, relative to character

the material gets more complex for most feature-character combinations. This means that the syntactical complexity in the German translation is higher than the one in the English original. It must be borne in mind that the scores being compared were first adjusted for the overall difference in typical syntactic complexity between the two languages (see Section Research Design). Therefore, a possible rise in complexity due to a generally higher syntactic complexity of the German language has been accounted for here and is not responsible for the higher scores of the translation. This is somewhat striking because the English subtitles are a verbatim transcript of the original dialogue, and therefore themselves already violate the recommended character frequency for subtitles of about 13 characters per second: a full, two-line subtitle with a maximum of 39 characters per line, i.e., 78 characters total, should be displayed for no more than six seconds according to industry standards (Díaz-Cintas/ Remael 2014: 84; 89). In the majority of cases, this is much less text than the equivalent of the spoken words of the original dialogue. The recommended character frequency of industry standards for subtitles is to enable viewers to be able to read the subtitles in time and follow the on-screen action at the same time. The German subtitles of the surveyed show are generally longer than the already too long English original ones, and even longer in relation to the German language than the English ones in relation to the English language. This increased text length is in part responsible for the increase in syntactic complexity. However, as has been said, this does not hold for all feature-character combinations. So, to answer the research question: although there is a noticeable upwards tendency as far as a change in syntactical complexity comparing original and translation is concerned, this change is not universal and syntactic complexity changes mainly depending on the respective combination of a syntactic feature and a character.

This means that the syntactic individuality of the original characters' ways of speaking has not been maintained because it changed to a different extent for each character and each syntactic feature. What can also be observed, is a general tendency of leveling syntactic complexity in the translated subtitles: from Table 2 can be seen that for two of the three syntactic features both the standard deviation and the coefficient of variation of the characters values for these features are lower compared to the original. So, stylistic individuality of characters, at least as far as syntax is concerned, is somewhat distorted in translation in two ways: it changes non-uniformly for the different characters, and it is generally leveled.

4. Discussion and Outlook

The first thing that needs to be discussed is a possible weakness of the presented approach. This is due to the fact that, first, the reference corpora used were web corpora, whose quality and representativeness for a given language is not beyond dispute. This, however, is true of any corpus and it would have gone beyond the scope of this survey to enter the discussion of corpus representativeness and the corpora were mainly chosen due to their size. Second, the software used for parsing, i.e., *UDPipe* (Wijffels 2018), is, as any parser, not perfect and a follow-up study could include a qualitative evaluation of its performance. The same is true of the packages used for syllabification.

Apart from these caveats, the gathered data and subsequent analyses raise a number of questions, which cannot be answered within the scope of this paper, but which are worth pointing out possibly going into in future work. One such interesting finding is the fact that the variation between characters (syntactic) speech patterns was not extremely marked, be it in the original or the translation, where it was even more leveled. The reasons for this can only be speculated about here. A likely explanation for this may be that the research material is scripted TV dialogue (and its representation in the form of subtitles) and therefore pseudo-natural speech. A leveling of syntactic patterns due to the simplification/shortening of utterances during the production of subtitles can be ruled out because, as has been stated, the original subtitles are verbatim transcripts and their German translations are even longer than that. It would be interesting to research if the syntactic complexity of completely natural speech differs more from person to person than that of scripted dialogue does. Pertinent data would be required in order to assess the significance of the observed variation in syntactic complexity between the fictional characters of the surveyed show. For this purpose, corpora of spoken language that has also been translated would be required. The only likely candidates for this would most likely be interpreting corpora. These, however, document a highly special mode of oral communication, i.e., mostly prepared

statements that are translated under special circumstances: those of interpreting. It would probably be more promising to use monolingual, but comparable corpora in order to find out about the syntactometric properties of natural spoken language, and even better, to empirically compile such corpora specifically for this purpose.

Another interesting question is whether the demonstrated lack of syntactic variability in scripted dialogue and even more so in its translation does in any way correlate with the enjoyment of a media product. Or, conversely, could an increase in syntactic variation, especially in translated subtitles, increase the enjoyment of a show? A pertinent study could establish whether syntactic complexity and its variation is an important stylistic feature as far as its bearing on the quality of original writing and its translation is concerned.

References

Baker, Mona (2000). "Towards a Methodology for Investigating the Style of a Literary Translator?" In: *Target* 12(2). Pp. 241–266.

Benoit, Kenneth/ Watanabe, Kohei/ Wang, Haiyan/ Nulty, Paul/ Obeng, Adam/ Müller, Stefan/ Matsuo, Akikata. (2018). "quanteda: An R package for the quantitative analysis of textual data". In: *Journal of Open Source Software* 3(30). Pp. 774–777.

Björnsson, Carl Hugo (1968). *Läsbarhet* (with an English summary). Stockholm.

Björnsson, Carl Hugo (1983). "Readability of newspapers in 11 languages". In: *Reading Research Quarterly* 18(2). Pp. 480–497.

Coşeriu, Eugenio (1981[1958, oral presentation]). "Los conceptos de 'dialecto', 'nivel' y 'estilo de lengua' y el sentido propio de la dialectologia". In: *Lingüística española actual*, III. Pp. 1–23. [n.v.]

Díaz-Cintas, Jorge/ Remael, Aline (2014). *Audiovisual Translation: Subtitling*. London/New York.

Faust, Manfred (1988). „Diaphasische Variation im Sprechen mit Ausländern". In: Albrecht, Jörn/ Lüdtke, Jens/ Thun, Harald (eds.). *Energeia und Ergon: sprachliche Variation, Sprachgeschichte, Sprachtypologie. Studia in honorem Eugenio Coseriu, vol. 2*. Tübingen. Pp. 501–510.

Fichtner, Edward G. (1981). „Measuring Syntactic Complexity: The Quantification of One Factor in Linguistic Difficulty". In: *Die Unterrichtspraxis/ Teaching German* 13(1). Pp. 67–75.

Flesch, Rudolph (1948). "A New Readability Yardstick". In: *Journal of Applied Psychology* 32(3). Pp. 221–233.

Goossens, Jan (1977). *Deutsche Dialektologie*. Berlin.

House, Juliane (1977). *A Model for Translation Quality Assessment*. Tübingen.

Huang, Libo (2015). *Style in Translation: a Corpus-Based Perspective*. Berlin.

Kenny, Dorothy (2001). *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester.

Lynch, Gerard (2017). "Strange bedfellows. Shifting paradigms in the corpus-based analyses of literary translations". In: *inTRAlinea. Online translation journal* 19. (www.intralinea.org/specials/article/2257, accessed 05.02.2020).

Majliš, Martin/ Žabokrtský, Zdenek (2012). "Language Richness of the Web". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (*LREC'12*). Paris. Pp. 2927–2934.

Olsson, Nikolaj Lynge (2019). Subtitle Edit. (www.nikse.dk/SubtitleEdit, accessed 24.03.19).

Rudis, Bob (2016). *Package 'hyphenatr'. Tools to Hyphenate Strings Using the 'Hunspell' Hyphenation Library*. (www.rdocumentation.org/packages/hyphenatr/versions/0.3.0, accessed 10.09.2018).

Rybicki, Jan. (2006). "Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations". In: *Literary and Linguistic Computing* 21(1). Pp. 91–103.

Rybicki, Jan (2012). "The great mystery of the (almost) invisible translator: Stylometry in translation". In: Oakes, M.P./ Ji, M. (eds.): *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam. Pp. 231–248.

Saldanha, Gabriela (2011). "Style of Translation: The Use of Source Language Words in Translations by Margaret Jull Costa and Peter Bush". In: Kruger, Alet/ Wallmach, Kim/ Munday, Jeremy (eds.): *Corpus Based Translation Studies: Research and Applications*. New York. Pp. 237–258.

Warner Bros. Entertainment Inc. (2012). *Two and a half Men, Staffeln 1–8 [Season 1–8]*. DVD box set.

Wijffels, Jan (2018). *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. [R package version 0.4]. (https://CRAN.R-project.org/package=udpipe, accessed 10.09.2018).

Winters, Marion. (2007). "F. Scott Fitzgerald's *Die Schönen und Verdammten*: A corpus-based study of loan words and code switches as features of translators' style". In: *Meta* 35(1). Pp. 412–425.

Andy Stauder
University of Innsbruck
Department of Translation Studies
Herzog-Siegmund-Ufer 15
A-6020 Innsbruck, Austria
csaf3004@gmail.com
ORCID: 0000–0002–2253–0614

Michael Ustaszewski
University of Innsbruck
Department of Translation Studies
Herzog-Siegmund-Ufer 15
A-6020 Innsbruck, Austria
michael.ustaszewski@uibk.ac.at
ORCID: 0000–0002–2000–5920