

Wrocław, 18.03.2024 r.

Prof. dr hab. Łukasz Grabowski
Instytut Językoznawstwa / Institute of Linguistics
Uniwersytet Opolski / University of Opole
Pl. Kopernika 11
45-040 Opole, Polska / Poland

Review of a doctoral dissertation by Judyta Mężyk, MA, titled “Pragmatemes in Audiovisual Translation in English-French-Polish Language Pairs” (327 pages)
supervised by Prof. Lucie Gournay (University Paris-Est Creteil) and Prof. Andrzej Łyda
(University of Silesia)

This review was prepared pursuant to a decision of the Scientific Board of the Institute of Linguistics (Faculty of Humanities) at the University of Silesia dated 20 February 2024.

Preliminary remarks

The doctoral dissertation under review makes an important contribution to research on formulaic language and (audio-visual) translation: it provides a detailed description of so-called pragmatemes, i.e. recurrent text chunks tied to particular social situations and speech acts, in terms of their formal-linguistic properties and their use in dialogues/subtitles of selected TV series available on the streaming platforms (Netflix). More precisely, the Author explored pragmatemes found in English, French and Polish subtitles, which were collected in a custom-designed study corpus, and this way obtained an insight into the pragmatemes' forms and textual functions in both native and translational texts in three languages. As regards translational texts, the Author also conducted a thorough quantitative and qualitative analysis of translation techniques applied to pragmatemes in English-French and English-Polish translation. Generally speaking, the Author correctly uses – adapting to the specificity of her research – selected quantitative and qualitative methods typical of corpus linguistics, combining them with theoretical concepts developed in the field of formulaic language and audiovisual translation. It is commendable that the Author undertook the difficult task of combining and analysing a huge amount of linguistic data from three typologically different languages.

Structure of the dissertation

The doctoral dissertation is 327 pages long and it consists of abstracts in English, French and Polish, Introduction, 4 chapters, Conclusions, extensive References and Appendices. The thesis can be divided into three different parts.

The first part, which includes Chapter 1, provides a detailed theoretical introduction to key concepts used in the study, which are later used to develop and operationalize the definition of a pragmateme, which is the unit of analysis used in the study. In particular, we find there a very detailed synthesis of relevant theoretical and methodological approaches to pragmatemes

grounded in Anglo-Saxon, Francophone and Slavic (notably Russian and Polish) research tradition. Also, an overview of the field of formulaic language research is presented in the chapter, followed by a synthetic introduction to selected aspects of audio-visual translation.

Chapter Two is devoted to methodology, where the Author describes the research material and the process of corpus compilation as well as the search for the units of analysis in individual languages (i.e. in the English, French and Polish sub-corpora of subtitles), which was based on the list of pre-selected items extracted from reference materials (primarily monolingual and bilingual dictionaries and phrasebooks), and in the parallel English-French and English-Polish sub-corpora of subtitles. All in all, the Author identified and later analysed 290 English, 186 French and 106 Polish pragmatemes.

In Chapter Three and Four, the Author presents the study results, starting with linguistic analysis of pragmatemes through the lens of selected formal-linguistic, distributional and pragmatic criteria, enhanced by statistical analyses (univariate and multivariate ones) conducted using primarily a test of statistical significance (chi-square test of independence) and correspondence analysis. This confirms that the Author is no stranger to exploratory data analytic techniques. This part is followed by a largely qualitative investigation of translation techniques (some of them fine-tuned by the Author who capitalized on earlier typologies popular in specialized literature) applied to the so-called *charged pragmatemes* in English-French and English-Polish audio-visual translation of subtitles.

The thesis ends with Conclusions, where the Author summarizes obtained research findings, discusses limitations of the study (in particular of the tools used) and outlines avenues for future research.

Below, I would like to discuss the strengths and weaknesses of the dissertation, notably with respect to corpus linguistic methods used in the study, which I am particularly interested in.

Originality and strenghts of the disseration

First, I would like to comment on and emphasize those aspects of the dissertation that contribute to its individual and creative character.

- This dissertation fills in the gap in empirical research on pragmatemes by providing a comprehensive and original description of their linguistic features, selected aspects of use (notably frequency) in everyday spoken language represented by TV series and film subtitles as well as their treatment in English-to-French and English-to-Polish translation. This makes the study both contrastive and translational, which means that pragmatemes are explored from the systemic and textual (language use in native and translational texts in each language) perspectives. The Author draws extensively from Anglo-Saxon, Francophone and Slavic (Polish and Russian) literature from the field of phraseology and formulaic language, which allows her to achieve a very ambitious goal set out in the dissertation, that is, to develop a comprehensive, universal (i.e. language-agnostic) definition of pragmatemes (p. 50), which is later operationalized for the purpose of empirical analyses. It is worthwhile emphasizing that the Author

capitalized on both widely-known (works by Mel'chuk, Burger, Blanco & Mejri, Kauffer, Kecskes, Wood, Wray, Cowie, Yorio, Nattinger & De Carrico, Coulmas, Vinogradov, Amosova etc.) and somewhat lesser known (Chlebda, Gębka-Moroz, Awdiejew etc.) approaches to phraseological and formulaic language research in Europe. This means that the Author knows a wide range of theories and methodologies used to study formulaic language, encompassing formal-linguistic, distributional, sociolinguistic and psycholinguistic ones, which was probably one of the reasons why a decision was made to treat both single-word and multi-word units (also exhibiting pattern variability) as pragmatemes in this study (in contrast to many earlier approaches favouring multi-word units, e.g. Blanco & Mejri 2018).

- The design and compilation of a tailor-made representative (given the scope of the study) corpus of English, French and Polish subtitles of contemporary TV series, including native and translational texts (translations from English). At this point, I would like to encourage the Author to make both the research material and all the data (including spreadsheets and an R script used for multiple correspondence analysis) available on data repository to ensure reproducibility and replicability of the study.
- A very clearly described methodology used in the study, notably the process of development of the initial/reference list of pragmatemes based on a wide variety of sources, followed by the application of the Unitex software to identify the pragmatemes in the study corpora (in particular, the use of Locate Pattern function helped increase the recall of the extracted data).
- Comprehensive and detailed analyses of the empirical material enhanced by data visualization that provided insights into linguistic, distributional and pragmatic (e.g. a very labour-intensive analysis of speech acts tied to pragmatemes, see p. 173) aspects of pragmatemes not only in native texts but also in English-to-French and English-to-Polish translation (see Chapter 4). Undoubtedly, the thesis makes an important contribution to the description of this rather underexplored unit of analysis, notably in the cross-linguistic and translational perspective, capitalizing on earlier cumulative knowledge.

Critical comments with respect to the doctoral dissertation

Below I have presented my critical comments with respect to some of the fragments of the thesis. They concern certain formulations, interpretations of specialized literature, methodological decisions etc. As such my critical remarks and polemics do not adversely impact my overall positive evaluation of the submitted dissertation.

Abstract and introduction

- The research questions (RQs) posed in the dissertation are not explicitly formulated (as questions per se): they are not included in the methodological section but they are stated in the abstract and introduction as either research goals or declarative statements (e.g. *this study's first research question aims to propose a universally applicable definition of pragmatemes; the second research question concerns the linguistic characteristics of pragmatemes; The third research question concerns pragmatemes*

in translation, specifically within one type of audiovisual translation, namely subtitling). All in all, as the study is primarily descriptive and combines both qualitative and quantitative methods of data analysis, the RQs, notably RQ2 and RQ3, should be more precise and specific in view of the chosen data analysis methods and quantifications.

Chapter 1

- The Author often explains that pragmatemes are key elements of language fluency, yet some examples provided in the thesis to illustrate this point represent linguistic errors or are somewhat artificial, i.e. they do not represent typical language use. For example (p. 4), has the Author found any evidence (in language corpora) of any competent language user (a native-speaker of English or a translator) saying *On your health* instead of *Bless you* (with the exception of uttering this phrase jokingly or as a manifestation of language use of a Polish learner of English as a foreign language). On p. 33, a phraseme *Smoking forbidden* is also rather unnatural in public places in English-speaking countries (in contrast to *No smoking*, *Thank you for no smoking* etc.). On a different note, I appreciate the Author's conscientiousness in that some glaring errors or untypical constructions (e.g. calques from Polish) were identified in printed reference materials (e.g. *A pleasant journey!* described on p. 134), and subsequently removed from the initial/reference list of pragmatemes used in the study.
- One of the goals of the thesis is to develop a universal (language agnostic) definition of pragmatemes as a unit of analysis. Such a definition is first presented on p. 50, where we can read that pragmatemes "are language units that are fixed, i.e. used without changes in their form" and – at the same time – "allowing for slots or not". Perhaps this should be clarified as it is essential in view of later operationalization of pragmatemes that takes into consideration pattern variability.
- On pp. 70-71, the Author comments on certain disadvantages of intuitive identification of pragmatemes based on a select set of criteria and notes that such a procedure is vulnerable to judgement inconsistency, also resulting from the fact formulaic language encompasses a continuum of different items, e.g. compositional and non-compositional ones, fixed and variable etc. Perhaps it should be mentioned that nowadays researchers make extensive use of inter-rater agreement metrics (cf. Brezina 2018) to ameliorate such problems.
- When reading an overview of formulaic language (Section 1.2) I have noticed that the Author occasionally relies too much on secondary sources (e.g. an excellent overview by Wood (2015)) rather than on primary ones presenting key concepts under discussion there. That is why some important aspects of those items, also in the context of pragmatemes, are missed, e.g. distributional and frequency thresholds in the case of lexical bundles, the concept originally proposed by Biber et al. (1999); for the same reason, we find a very general definition of metaphors proposed by Wood (2015); an inadequacy in that the concept of concgrams was first proposed in 2006 by Cheng et al. rather than in 2009 etc.

Chapter 2

- On p. 123, I find a fragment that is not clear to me: “To study the translations, it would be preferable for the corpora to be not just parallel corpora, but rather corpora with actual translations of the source texts.”. To my knowledge, parallel corpora (e.g. *InterCorp*, *Paralela*) include source-texts paired (aligned) with their translations, so the phrasing “actual translations” may not be clear to readers. Also, I do not agree with the claim found on the same page that spoken corpora do not necessarily demonstrate how pragmatemes are used in everyday conversations because they contain mostly quasi-spoken language (transcripts of interviews etc.). However, corpora such as Spokes (Polish conversational corpus, cf. Pezik 2015) or BNC Spoken 2014 include transcripts of conversations recorded in informal settings (typically at home), among friends and family members and thus represent naturally-occurring language in everyday situations. This makes them suitable sources of research material to study the use of pragmatemes.
- On p. 138, the Author discusses the use of regular expressions in the Locate Pattern search window of Unitex software, but the provided example “<A> <N>” represents morphological tags. Does the software have in-built morphological annotation? Were the corpora used in the study morphologically tagged (using taggers for English, French and Polish)?
- On p. 142 and elsewhere, I found figures (e.g. Figure 12) that are unreadable on printed page. I strongly recommend that smaller data excerpts (e.g. top-5 rows) be presented in any future publications (should the partial results or entire work be published).

Chapter 3

- On p. 149 and elsewhere, we occasionally find fragments that do not contribute much to the discussion or repeat information presented earlier in the text, e.g. *The present study endeavors to further enhance the description of pragmatemes, asserting that in order to investigate a phenomenon effectively, one must possess a precise comprehension of the subject.*
- On pp. 190-195, the Author presents an experiment in using chi-squared test for independence (described in greater detail in Navarro 2015, Levshina 2015, Brezina 2018, among others), but I find certain fragments imprecise and incomplete. The question tested there is in fact whether the categorical variables Language (a factor with three levels: English, French and Polish) and Construction Type (a factor with two levels: e.g. elliptical or non-elliptical, which can be treated as a binary variable) are independent or dependent (associated). The problem is that the data¹ represents three distinct sub-corpora of three typologically-different languages² (the data are

¹ NB! It is customary that expected values be rounded to the nearest whole number.

² In addition: since the data was sampled from different populations, we note considerable differences between the numbers of pragmatemes (286 in English, 186 in French and only 106 in Polish). That is why one needs to exercise caution when comparing statistical metrics (in a similar way that the Author correctly did

sampled from three populations rather than from a single population, which is – fortunately – the conclusion that the Author arrived at on p. 194³. If instead of Language the Author had used Variety (e.g. genre of TV series etc.) and conducted the test separately for each language (subcorpus), then the results would be interpretable. It would have been clearer if the Author had explained how the test statistic (X^2) is calculated; also, the test statistic and degrees of freedom are not provided and we only get p-values. To make matters worse, the Author also did not calculate effect size measures (typically odds ratio or Cramer's V for categorical variables, which are the types of variables for which chi-square test of independence is used). Calculation of effect size pertains to the problem that the Author mentioned (and misinterpreted a bit) on p. 195, namely that apart from checking whether the association is statistically significant (=whether variables are independent or not) it is necessary to measure the strength of the association (or, in other words, its magnitude), which helps us assess whether the finding is linguistically meaningful or important in practice. It is also important to remember that the pragmatemes explored in this study are not extracted directly from the three sub-corpora but they are identified in a top-down approach using the initial reference list and a very general definition proposed in this thesis. In reality, we do not precisely know how many pragmatemes are found in each subcorpus. Had the initial reference list (and the identification criteria) been different, the numbers of pragmatemes (both types and tokens) would have been different as well.

- Overall, the statistical analyses leave the reviewer a bit unsatisfied in that the Author did not employ multifactorial statistics, especially mixed effect models (cf. Winter 2019, Gries 2021). The Author explained clearly in the thesis that the use of pragmatemes is a multi-faceted phenomenon (notably in audiovisual translation, where it is subject to additional constraints, e.g. time and space), so using mixed effect models would allow to test not only descriptive hypotheses but also explanatory ones, and generalize from the obtained findings. Much as I realize that it was beyond the scope of this dissertation, I cordially hope that the Author would consider such research avenue in the future. The phenomena studied in the thesis are indeed multifactorial in nature.

Chapter 4

- As for the analysis of the frequency of use of particular translation techniques, I wonder whether there have been any cases of problematic assignment to one rather than another technique?
- On p. 250, the Author writes about the concept of equivalence in a very simplified manner, without referring to any particular approach and definition of this key concept (e.g. symmetrical vs asymmetrical, directional vs natural, textual vs systemic, formal

when selecting data visualization methods, notably mosaic plots, which take into consideration raw numbers and proportions).

³ This also applies to multiple correspondence analysis (pp. 199-206), where the Author first conflated the data for three language, but fortunately later conducted MCAs separately for each language.

vs dynamic, full vs partial vs zero etc.). This is a weak point of the thesis. Surely, equivalence is tackled differently from the perspective of language system or language in use, and I agree with the Author – following the famous claim made by Pym (2010) – that equivalence is a simple idea yet it is difficult to apply.

- The Author devoted so much space in the dissertation to explaining the complexity and specificity of both audio-visual translation (subject to various specific constraints) and pragmatemes so it is not clear to me why an attempt at verification of target language equivalents was undertaken using a neural machine translation system, which processes linguistic items in isolation? In fact, pragmatemes in source texts (subtitles) are used in a given situational context so the benefit (setting aside reliability and representativeness of the results abstracted away from context) of bidirectional verification – using NMT engine – of the pragmatemes' equivalents found in Polish and French translations of English subtitles may not be entirely clear to readers. Perhaps this should be explained to readers.

Conclusions

- The Conclusions would benefit from a more detailed discussion of how the obtained results could be applied, e.g. in translation teaching or translation practice, notably in the case of audio-visual translation. Also, as this study is primarily descriptive, it would be also interesting to comment upon avenues for future research – most probably conducted using multifactorial statistics – targeted at the very explanation of observed linguistic patterns and tendencies in translation.

Assessment of language and editing

The linguistic and editing aspects of the reviewed dissertation deserve a high rating as the Author uses mature academic English. The dissertation is also perfectly prepared in terms of editing and data visualization. In the course of the review, I identified some rare grammatical and punctuation errors, e.g. in the abstract: *depends on the context, research questions of this study, primary focus in on the context* etc.; p. 40 - i.e., *its reproducibility in a given situation and to express a given message*; p. 122 – *Bruti notes that recently, the distinction has become ...* ; p. 123 – *in the these language*; p. 124 – *in a fun way*. I would recommend that the thesis (notably the abstracts in English and Polish) be carefully proofread by native speakers of these languages.

Conclusions and final decision

The doctoral dissertation submitted by Judyta Mężyk, MA represents a mature study combining quantitative and qualitative methods used to explore linguistic features of pragmatemes found in native texts and audiovisual translations. The very design and compilation of the comparable-parallel corpus of English, French and Polish subtitles used as a research material is an additional scientific achievement and it represents a value in itself: as a publicly available dataset (if possible) it may come in particularly useful for researchers in the future. Irrespective of certain critical comments that I have expressed in this review, which are aimed to increase the Author's awareness of certain theoretical and methodological

aspects, my overall evaluation of the dissertation is positive. That is why I hereby state that the thesis titled “Pragmatemes in Audiovisual Translation in English-French-Polish Language Pairs” is sufficient to be accepted as a doctoral thesis. The dissertation meets the requirements specified in the applicable legal provisions in Poland. Thus, I expressly request that Ms Judyta Mężyk, MA, be admitted to further stages of the doctoral procedure.

/prof. dr hab. Łukasz Grabowski/



Bibliography⁴

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Cheng., W, Greaves, C. & Warren, M. (2006). “From n-gram to skipgram to concgrams”. *International Journal of Corpus Linguistics*, 11 (4), 411-433.

De Sutter, G. & Lefer, M.-A. (2019). “On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach”. *Perspectives*, 28:1, pp. 1–23.

Gries, S. (2021). (Generalized Linear) Mixed-Effects Modeling: A Learner Corpus Example. *Language Learning*, 71(3), 757-798.

Levshina N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Navarro D. (2015). *Learning Statistics with R: A tutorial for psychology students and other beginners*. (Version 0.6) Sydney: University of New South Wales.

Pęzik, P. (2015). ”Spokes – a Search and Exploration Service for Conversational Corpus Data”. In *Selected Papers from the CLARIN 2014 Conference*, October 24-25, 2014, Soesterberg, The Netherlands, 99–109.

Pym, A. (2010). *Exploring Translation Theories*. London and New. York: Routledge.

Winter B. (2019). *Statistics for Linguists: An Introduction Using R*. London: Routledge.

⁴ I skipped those references found in the dissertation.