https://doi.org/10.18485/belida.2022.1.ch14

Milica Dinić Marinković*
University of Belgrade – Faculty of Philology

(SOME SIMPLE) MEANS FOR EXTRACTION OF ORTHOGRAPHICALLY UNMARKED FICTIONAL DIALOGUE

Abstract

The main aim of this research paper is to investigate whether it is possible to extract orthographically and typographically unmarked fictional dialogue from unannotated corpora with a simple regex query, and to what extent. In order to give an answer to this question, the paper provides a brief review of recent approaches and proposed solutions to this problem, with special attention on work with unannotated corpora. The simple means for extraction proposed here are based on the identification of overt linguistic features of character strings in orthographically and typographically unmarked fictional dialogues, and different ways of their conversion into patterns for the extraction from unannotated corpora. As a research corpus we used Sally Rooney's novel "Normal people", both the original version (English) and Serbian translation.

Keywords: unmarked fictional dialogue, unmarked direct speech representation, overt linguistic features of a fictional dialogue, extraction patterns from unannotated corpora, Sally Rooney's fictional dialogue

* milica.dinic.marinkovic@fil.bg.ac.rs

1. Introduction

One of the simple exercises for students in any corpus linguistics course is to form a regular expression[1] that will match direct speech strings/fictional dialogue in a corpus of prose fiction. This is a pretty convenient task because the solution to this problem lies in the identification and defining of orthographical and typographical conventions for direct speech representation in a given corpus and does not require corpus annotation.

Although typography of fictional dialogue markers varies in regard to specific convention – publishing style, "fashion" of the epoch, as well as regional or national tradition, in these diachronic and/or diatopic variations we could easily find the most frequent typographical markers of direct speech. For SerBoCroation publishing tradition these markers are quotation marks and long dashes, so, according to this, there are two simple types of regular expressions for direct speech strings matching. In the case of quotation marks, one could use the following simple regular expressions.

E.g. (1) „.+**?"**

E.g. (2) „[**A-Za-z**[2]]+ [**0-9**\.,**?!**]*"

But the question is what happens when orthographical and typographical markers of a fictional dialogue in the text are missing?

[1] *Regular expression*, or *regex*, is a pattern that consists of a sequence of characters, and it is used for formal description of character strings, or sequences of character strings in a text. The main usage of the regex is a specification of search patterns in some corpus/text, and all the character strings that correspond to the query in form of a regex are called *matching strings*. (For more about regex see Jurafsky & Martin, 2021).

[2] This range of characters corresponds to ASCII character encoding. For the Serbian language there are two different ranges of characters, depending on the alphabet, i. e. character code which is being used in the text – Cyrillic or Latin.

1.1. Background, research questions, and goals

We are witnessing the trend of intentionally neglecting traditional punctuation conventions in contemporary literature (e. g. Sally Rooney, Bernardine Evaristo, Milena Marković), among which Sally Rooney's novel *Normal People* is undoubtedly significant as a globally popular phenomenon[3]. According to its status, this novel could easily be a corpus for students' exercises. Yet, since this novel lacks quotation marks in fictional dialogues (as do other novels by the same author), it is impossible for students to fulfill the seemingly simple task of matching direct speech strings in this text.

This particular challenge caught my attention during the course of reading the novel *Normal people* when I realized I had not been noticing the lack of quotation marks for about 20 pages, although my perception of dialogues had not been impaired by that. This experience confirms several things.

Firstly, this partly confirms the stance of Cormac McCarthy (the author who also doesn't use quotation marks in his writing) that:

> "*There's no reason to blot the page up with weird little marks. I mean, if you write properly you shouldn't have to punctuate.*"

Sally Rooney herself stated:

> "*I can't remember ever really using quotation marks. I didn't see any need for them, and I don't understand the function they perform in a novel.*"[4]

Besides the clarity of writing, viewing this matter from the psycholinguistic perspective implies that the linguistic characteristics of a fictional dialogue have a clear representation in our mind, a

[3] Although various sorts of planned neglect of the different orthographical and typographical conventions have been presented in the history of literature, particularly in the epoch of Russian Futurism, it has not been the mainstream. There are only several, mostly Irish writers, who keep minimalist usage of punctuation in order to achieve maximal clarity in writing, e. g. James Joyce, Samuel Beckett, and Roddy Doyle.

[4] https://www.palatinate.org.uk/the-end-of-conventional-punctuation/

representation that does not depend on typographical markers of fictional dialogues but on features of a different kind.

However, these observations of novelists are significant for the investigation and findings on the nature and characteristics of written language. According to the literature on features of written language (Halliday, 1989; Halliday & Martin, 1993), it is doubtless that the differences between typical oral and typical written language are fundamental, not only regarding their form but in their function and content as well. Of course, like in every social phenomenon, between these prototypical instances of written and oral forms of language use, there is a whole spectrum of various sorts of mixed or hybrid forms. The literary form of novel surely belongs to a hybrid form of written language, because it contains instances of constructed oral communication represented in fictional dialogues.

As Abercrombie stated, "(...) *the aim of writing is not, usually, to represent actual spoken utterances which have occurred*"[5]. When Halliday examines which features of oral language are represented in written form, he concludes:

> "*So the omission of prosodic features from written language is, in some respects and under certain circumstances, a genuine deficiency. There is, on the other hand, a device that is used in order partially to overcome this deficiency; this is the device of punctuation.*" (1989: 32)

Further, by analyzing the functions of punctuation, Halliday concluded that the main purpose of punctuation is not to represent some of the prosodic features of oral language, but that it evolved within the written language for the purpose of achieving greater clarity of the written text itself (*Ibid.*: 32–39). The history of emerging and conventionalization of quotation markers confirms this standpoint[6].

Bearing in mind the writers' statements, as well as the finding of research on written language, one could anticipate that in the future the number of writers who avoid conventional punctuation may rise, while corpus and computational methods for extraction are based

[5] Abercrombie, 1965, as cited in Halliday, 1989: 31.

[6] On the history of punctuation markers see Houston 2013.

either on punctuation characteristics of character strings or on very sophisticated tools and time-consuming construction of resources.

Therefore, the research questions of this paper are: (1) what would be the simple means for the extraction of an orthographically and typographically unmarked fictional dialogue from unannotated corpora, and (2) whether it is possible to extract unmarked fictional dialogues from unannotated corpora by simple regex (almost) matching, and to what extent?

In order to answer these questions, the main aims of this paper are (1) identification of linguistic features of strings in orthographically and typographically unmarked fictional dialogue, and (2) construction of patterns for their extraction from unannotated corpora.

As a research corpus we used Sally Rooney's novel "Normal people", both the original version (English) and Serbian translation[7].

2. *Fictional dialogue* in recent linguistic research

Fictional dialogue is usually defined as "passages of character-character conversation within a literary text" (Nykänen & Koivisto, 2016: 1), or „passages of characters' direct speech in prose fiction text" (Kurfalı & Wirén, 2020: 105). Cited definitions represent a narrow understanding of this narration mode. Broader, and, in our opinion, more precise determination of fictional dialogue includes all the characters' interactions which stand opposite to the narrator's telling the story[8]. That includes, for example, an e-mail correspondence of characters', their SMS or chat conversation, and also internal dialogue.

Besides the term fictional dialogue, the following terms are being used (though less frequently): *imaginary dialogue* (Kinzel & Mildorf,

[7] Only for the purpose of the research presented in this paper .txt form of published books have been made. In further processing of text, AntConc software was used. For details on the analyzed corpus and used tools, see the citations at the end of this paper.

[8] As in any other phenomenon, there are all sorts of different manners of characters' interaction representation, e. g. „free indirect style, *style indirect*

2012), *constructed dialogue*, *hypothetical dialogue* (Semino et. al., 1999), *characters' discourse* (Ek & Wirren 2019).

The phenomenon of a fictional dialogue occupies the attention of researchers in various linguistic fields – Stylistics and Corpus Stylistics (Toolan, 1985; 1987; Leech & Short, [1981], 2007; Oostdijk, 1990; Semino & Short, 2004; Axelson, 2009; Ковачевић, 2013; Mahlberg et. al., 2019), Conversation Analysis (Bahtin, 1980; Burton, 1980; Thomas, 2012), Representology (Ковачевић, 2015), Literary Linguistics (Nykänen & Koivisto, 2016), Natural Language Processing (Byszuk et. al., 2020; Wirén, Ek & Kasaty, 2020; Kurfalı & Wirén, 2020; Ek & Wirén, 2019; Weiser & Watrin, 2012; Yeung & Lee, 2016 & 2017; Axelson, 2009; Oostdijk, 1990).

The growing interest in fictional dialogue within the field of natural language processing is driven by the need for automatic extraction of fictional dialogue for the purpose of providing a large amount of empirical data for systematic linguistic analyses. There are various research questions that depend on systematic collections of linguistic data of this kind, such as (1) questions of narrative structure, which include: distinguishing narration and speech, keeping track of addresses, identification, and modeling of fictional characters; (2) stylistic characteristics of fictional dialogue; (3) conversation analysis key questions concerned with the relation, similarities, and differences between real-life conversation and constructed conversation represented in fictional dialogue, (4) getting data on and analysis of (written) speech-like language from historical periods[9], and (5) questions about the very nature of fictional dialogue, which include description, determination, and positioning in regards to both spoken conversation and other narrative means in fiction.

libre, represented speech and thought, quasi-direct discourse, and combined discourse" (Toolan, 2001: 119), „free indirect speech" (Oostdijk, 1990: 236), but these phenomena are out of reach of the research presented in this paper. For terminological apparatus for different representations of speech in Serbian see Ковачевић, 2013 and Ковачевић, 2020.

[9] About challenges in historical pragmatics regarding the issue of collecting suitable data see Lalić, A. *Surpassing the "bad data" problem: Italian epistolary discourse as a source of spoken language* in this volume.

3. Some recent approaches to fictional dialogue automatic extraction

As Axelsson (2009) accurately distinguished, challenges for automatic extraction of fictional dialogue lie upon numerous linguistic, both theoretically fundamental and practical, unsolved problems. The first problem that Axelsson detected is the very definition of fictional dialogue, as its scope may vary in regard to concrete research questions (2009: 191). The other two major issues that she addressed originate from the treatment of prose fiction in referent corpora (BNC 2001, 2007), both in terms of representativeness of the samples that were included, and in terms of lack of direct speech annotation/tagging.

Approaches to the problem of automatic extraction of orthographically and typographically unmarked fictional dialogue could be systematized in the following way.

In most cases, the starting point is the process of annotation, but the scope and levels of annotation vary greatly. Some researchers use lemmatized corpora with in-depth MSD[10] annotation[11] (Ek & Wirén, 2019;), as well as different levels of semantic annotation (Wirén *et. al.*, 2020). This annotation process could be manual or semi-automatic. Others conduct manual annotation of targeted markers, e. g. speech framing verbs, sentence type, and/or structural elements (Oostdijk, 1990), or carry out manual annotation of the macrostructure of text (Ek & Wirén, 2019; Wirén *et. al.*, 2020).

A central step in recent approaches to fictional dialogue automatic extraction is a machine learning process, i.e. machine training. These processes are always, in the first instance, based on typographically unambiguously marked direct speech, which serves as material for machine training. In the next step, some researchers remove typographical markers (Kurfalı & Wirén, 2020), or use a multilingual approach and deep learning with *Bidirectional Encoder*

[10] MSD stands for morpho-syntactic descriptions.

[11] Although extraction could be accomplished regardless of MSD tagging, MSD annotation is done due to further linguistic analysis.

Representations from Transformers (Devlin *et al.*, 2018) so that machine classifiers do not rely on typographic markers (Bysuzuk *et. al.*, 2020).

Contrary to dominantly machine learning approaches, there is a more linguistic approach for automatic extraction of direct speech. This method proposed the creation of a formal unlexicalised grammar of direct speech representations and the automatic construction of an e-dictionary of lexical elements that can introduce direct speech instances. In the study of Weiser & Watrin (2012) unmarked direct speech instances are described as syntactic structure, and finite state automata were built (in Unitex software) according to recognized patterns of structure. E-dictionary, which consists of speech framing verbs, was used as direct speech "hunter".

4. Analysis part 1: Overt linguistic features of unmarked fictional dialogue

In work with unannotated corpora it is impossible to conduct an analysis on a word level. Hence the investigation of overt linguistic features of unmarked fictional dialogue in our research has been done on the sentence level. During the analysis of overt linguistic features of strings in instances of fictional dialogue in the analyzed corpus, and according to solutions proposed by Weiser & Watrin (2012)[12], the following components emerged as significant: (1) structural elements of an unmarked fictional dialogue, (2) order of structural elements in instances of unmarked fictional dialogue, (3) lexico-grammatical features of extracted instances, and (4) their sentence punctuation.

4.1. Structural elements of an unmarked fictional dialogue that are singled out are:

- unmarked direct speech (UDS)
- comma (,)
- speech framing verb (SFV)
- reference to the character – personal name or pronoun (RTC)

[12] With appropriate adjustments.

It is important to emphasize that, since the research question of this paper is whether it is possible to extract unmarked fictional dialogue from unannotated corpora, we could only lean on instances of dialogue already marked by the writer/narrator. This means that analyzed instances of fictional dialogue do not fit the narrow concept of fictional dialogue, which excludes instances of the narrator's framing of characters' conversation from the fictional dialogue, together with references to the characters.

4.2. The (linear) order of structural elements in unmarked fictional dialogue

[And I hear you did very well]	[,]	[she]	[says].	
1.	**2.**	**3.**	**4.**	(NP[13])
UDS	**,**	**RTC**	**SFV**	
[A čujem i da si odlično uradio]	[,]	[nastavila je]	[Lorejn].	
1.	**2.**	**3.**	**4.**	(NLj)
UDS	**,**	**SFV**	**RTC**	
[He was top of the class]	[,]	[says]	[Marianne].	
1.	**2.**	**3.**	**4.**	(NP)
UDS	**,**	**SFV**	**RTC**	
[Najbolje u celom razredu]	[,]	[dodala je]	[Merijen].	
1.	**2.**	**3.**	**4.**	(NLj)
UDS	**,**	**SFV**	**RTC**	

In the Serbian part of the analyzed corpus the order of structural elements is fixed – [UDS] [,] [SFV] [RTC], while in the English part of the corpus we find an alternating order of the last two elements – RTC and SFV, which can occupy third or fourth position in this type of fictional dialogue. The reason for the inalterable order of structural elements of fictional dialogue in the Serbian part of the corpus is the specific canonical form of speech framing verbs in fictional dialogue (see section 4.3), which does not allow a change in word order, and therefore in order of structural elements.

[13] Abbreviation for analyzed corpus: NP stands for the English version (*Normal People*), NLj stands for the Serbian version (*Normalni ljudi*).

4.3. Lexico-grammatical features of some structural elements

On the lexico-grammatical level, the least variable aspect of fictional dialogue instances represents the narrator's framing of the characters' conversation, which is achieved by using *speech framing verbs*, *lexical quotatives*, and *quotative markers* (Panić Cerovski & Ivanović, 2016: 143), or, traditionally *verba dicendi*.

These words do not just form a specific paradigmatic lexical set, but they also appear in a specific canonical form within the fictional dialogue (at least this is the case with the analyzed corpus).

In the English version of the novel, speech framing verbs appear in 3rd person Sg simple present tense more frequently than in the past tense (e. g. „to say" – says: freq. 751, said: freq. 480). In the Serbian version, speech framing verbs appear only in one form – 3rd person Sg past tense inverted form, e. g. *rekao je* (=he said), *upitao je* (=he asked), *iznenadila se* (=she got surprised), *nastavio je* (=he continued). This is a specific canonical form, determined by the initial position in the embeded close, that governs the position of reference to the character.

4.4. Sentence punctuation

When we take into account the characters of strings, it is impossible to exclude punctuation from consideration. Since it is clear that the full stop could not be considered a marker of any kind in the phenomenon we investigate here, we have analyzed the usage of exclamation and question marks in our corpus.

Conducted analyses show that the exclamation marks exclusively appear within a fictional dialogue, or at the boundaries of direct speech instances, which was expected. Indeed, it would be very odd for the writer to yell at the readers. Humor aside, it is not surprising, since there is no place for imperative or exclamatory sentences in the narration – direct commands, requests, warnings, or expressing of strong emotions, or other speech acts that these sentences represent.

Apart from the exclamation marks, question marks also belong exclusively to the strings of fictional dialogue in the analyzed corpus. There are two types of (quasi) exceptions that have been recognized. These are:

(1) occurrences of „?" in email texts that the female protagonist reads, and
(2) occurrences of „?" in the protagonists' internal dialogue.

According to the narrow concept of fictional dialogue – as the passages of characters' conversation within the text of prose fiction, these instances should be excluded. But, as we stated in section 2, we consider that everything which is not the writer's telling of the story belongs to a fictional dialogue. In addition, these two types of context – email text, and internal dialogue, belong to the mode of character interaction, not to the narration.

Although exclamation and question marks exclusively appear within a fictional dialogue, or at the boundaries of direct speech instances, most instances of fictional dialogue in the analyzed corpus do not contain these punctuation marks. Thus, even though the patterns for the extraction containing these marks will not match something other than fictional dialogue, the majority of the instances of fictional dialogue will stay unrecognized.

5. Analysis part 2: Patterns (from features) for extraction

Formal linguistic features that are presented and described in the previous section are converted into patterns, i. e. regular expressions for the matching and extraction of fictional dialogue in the analyzed corpus.

The first regex that was tested is based on structural elements and their order in instances of fictional dialogue (see 4.1 and 4.2).

Regex (1): ^([A-Z])\w+ [ ]* [A-Za-z]+, [A-Za-z]? (says|asks|...) [A-Za-z]?

This regex corresponds to the following pattern.

[new line] [word start with capital]
[word] {0-as many as there are words in a sentence} [word] [,]
[reference to a character] {0,1}
[sequence of speech framing verbs in suitable form in disjunctive relation]
[reference to a character] {0,1}

Thus, every instance of fictional dialogue (those that belong to the representation of the conversation between characters in a novel) always starts with a new line followed by a word starting with a capital. Between the initial word in a passage, and an obligatory element – a word followed by a comma, it could be null or many appearances of words (special character * stands 0 or more of them). After the word followed by a comma, there comes the element which is not obligatory in that precise position (in English), and that is why the special character "?" stands in Regex (1). This structural element (a reference to a character) could take the position (slot) before or after the speech framing verb, but the presence of both speech framing verbs and references to the characters are obligatory structural elements in this pattern of fictional dialogue instances.

Regex (1) matches, for example, the following instances of fictional dialogue:

The lads are fairly late, says Lisa.
Momci bogami baš kasne, prokomentarisala je Liza.

If they don't show up I will actually murder Connell, says Rachel.
Ako se ne pojave, ima da ubijem Konela, rekla je Rejčel.

Da li ikada sretneš Pegi u Dablinu? upitala ga je.
Do you ever see Peggy in Dublin? she says.

Jesi dobro? upitala ga je Lorejn.
Are you alright? says Lorraine.

Regex (1) matches only those passages of typographically unmarked fictional dialogue which contain the narrator's interference, i. e. framing. In the analyzed corpus instances that contain speech

framing verbs appear in about every 3rd – 5th paragraph of fictional dialogue, or maybe it is more precise to say – in about every 3rd – 5th turn within the characters' discourse, as Rooney uses paragraph breaks to indicate turns in conversation (see excerpts in text boxes below).

Text Box 1: Excerpt from the NP

(**1.p**) She shut her eyes. I do like you, **she said.**
(2.p) Well, if you met someone else you liked more, I'd be pissed off, okay? Since you ask about it. I wouldn't be happy. Alright?
(3.p) Your friend Eric called me flat-chested today in front of everyone.
(**4.p**) Connell paused. She felt his breathing. I didn't hear that, **he said.**
(5.p) You were in the bathroom or somewhere. He said I looked like an ironing board.
(6.p) Fuck's sake, he's such a prick. Is that why you're in a bad mood?
(7.p) She shrugged. Connell put his arms around her belly.
(**8.p**) He's only trying to get on your nerves, **he said.** If he thought he had the slightest chance with you, he would be talking very differently. He just thinks you look down on him.

Text Box 2: Excerpt from the NLj

(**1.p**) Zažmurila je. Pa sviđaš mi se, **rekla je.**
(2.p) A ja bih popizdeo kad bi upoznala nekoga ko bi ti se više dopao. Eto to bi bilo. Kad si me već pitala. Ne bi mi uopšte bilo svejedno. Da znaš. Okej?
(3.p) Tvoj drugar Erik mi je pred svima rekao da sam ravna ko daska.
(**4.p**) Nije joj odmah odgovorio. Osećala je njegov dah. Ja to nisam čuo, **rekao je**.
(5.p) Bio si u kupatilu ili šta znam gde. Rekao je da izgledam kao daska za peglanje.
(6.p) On je drkadžija, jebo te. Jesi zbog toga loše raspoložena?
(7.p) Slegnula je ramenima. Obgrlio ju je oko struka.
(**8.p**) On samo hoće da te iznervira, **rekao je.** Kad bi kapirao da ima i najmanje šanse kod tebe, pričao bi sasvim drugu priču. On samo misli da ga potcenjuješ.

Thus, the majority of the instances of fictional dialogue cannot be recognized. On the other hand, we think that tracking of writer's interference in the fictional dialogue, and catching patterns of it could give valuable results. It seems that the phenomenon in question, besides being driven by textual cohesion (see Polovina, 1999), is of

a discourse-pragmatic nature similar to one that takes place during the course of conversation (Polovina, 1988; Savić & Polovina, 1989), and especially as it is described in the chapter *Non-omission of deictic personal pronouns* in this volume (Polovina, 2022).

It is clear that the presented regex is too long and that it does not meet the condition of minimalism and elegancy, nor is it possible to form a regex of adequate size according to all the conditions that should be defined by it. Besides that, this regex also matches the strings which are not instances of fictional dialogue, but they appear on the boundaries of it, such as:

> *He waits for the coughing to subside, and then says*[14]*: What does he do to you?*
> *Sačekavši da mu prođe napad kašalj, rekao je: A šta ti on to radi?*

The solution for these particular cases for the English version is:

Regex (2): "(says|asks...):", or the following pattern:

[sequence of speech framing verbs in suitable form in disjunctive relation] [:]

This implies, as do the other examples too, that the solution to the problem lies in the combining of structural elements into the smaller distinct (sub)sets, instead of constructing one (long) regex with the potential to cover as many cases as possible.

6. Conclusion: The reach/limits of simple means

The main question of the research presented in this paper is whether it is possible to extract unmarked fictional dialogue from unannotated corpora by simple regex (almost) matching, and to which extent. In order to give the answer to this question identification of overt linguistic features of strings in orthographically and typographically

[14] The underlined part of the sentence is the one that is matched by the Regex (1).

unmarked fictional dialogue, as well as the construction of patterns for their extraction from unannotated corpora has been done.

It emerged that the main problem of unmarked fictional dialogue matching is the inability to define the boundaries of a fictional dialogue. Because of this, we cannot use a regex for matching the exact and whole strings of fictional dialogue, nor is it possible to match (and extract) all the instances of fictional dialogue in the analyzed corpus. The reasons for this limitation lie in two areas – in the area of conceptual determination of a fictional dialogue, and in the area of linguistic description for the purpose of natural language processing. From the conceptual point of view, the boundaries of a fictional dialogue are not invariably defined. At one end of a continuum, there are researchers who take into account only instances of conversational dialogue, while on the other end of the continuum, there are researchers who consider that all the instances which are not direct writer's words, i. e. narration belong to a fictional dialogue. As far as the formal description of boundaries of a fictional dialogue is concerned, one could be able to define the end of certain types of fictional dialogue, but we are not able to define the beginning of an unmarked direct speech passage in an unannotated corpus/text.

In addition to the incapability of defining boundaries of an unmarked fictional dialogue, there is yet another limitation. The conducted analyses presented in this paper show that in a work with unannotated corpora it is possible to detect only those passages of a typographically unmarked fictional dialogue which contain narrator interference, i. e. framing. And one should bear in mind that every 3rd–5th paragraph of fictional dialogue contains speech framing verbs.

Also, there is a challenge with a very long and robust form of regular expressions that have the potential to match more instances of fictional dialogue. This particular challenge could be overcome by dividing and organizing patterns into smaller distinct subsets. This means that the procedure of matching and extracting fictional dialogue instances by using simple regex queries needs to be repeated several times with different smaller regexes. Of course, not all instances of direct speech would be recognized, but they are not recognized with

more sophisticated means either. However, this type of task is well suited for students' exercises in corpus linguistics, and there is a lot of material for the practicing of regex formation.

References

Axelsson, K. (2009). Research on fiction dialogue: Problems and possible solutions. In: Jucker, AH, Schreier, D, Hundt, M. (Eds.), *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, (189–201).

Bahtin, M. (1980). *Marksizam i filozofija jezika*. Beograd: Nolit.

Burton, D. (1980). *Dialogue and Discourse. A Sociolinguistic Approach to Modern Dialogue and Naturally Occurring Conversation*. London: Routledge and Kegan Paul.

Byszuk, J., Wozniak, M., Kestemont, M., Lesniak, A., Łukasik, W., Šela, A. & Eder M. (2020). Detecting direct speech in multilingual collection of 19th-century novels. In: *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, European Language Resources Association (ELRA), 100–104.

Ek, A. & Wirén, M. (2019). Distinguishing narration and speech in prose fiction dialogues. In: *Digital Humanities in the Nordic Countries 4th Conference (DHN)*, Copenhagen, Denmark, 124–132.

Halliday, M., A., K. (1989). *Spoken and Written language*. Oxford University Press.

Halliday, M., A. K. & Martin, J. R. (ur.), (1996 [1993]). Writing Science, Literary and Discourse Power. London: The Falmer Press.

Houston, K. (2013). *Shady characters: the secret life of punctuation, symbols, & other typographical marks*. First edition. New York – London: W.W. Norton & Company.

Jurafsky, D. & Martin, J. H. (2021). *Speech and Language Processing* (3rd edition – draft). Available at https://web.stanford.edu/~jurafsky/slp3/2.pdf

Kinzel, T., & Mildorf, J. (Eds.) (2012). *Imaginary Dialogues in English: Explorations in a Literary Form*. Heidelberg: Universitätsverlag Winter, 9–30.

Ковачевић, М. (2013). О терминосистему туђег говора, In: *Лингвистика као србистика*, Vol. 1, (pp. 67–99). Универзитет у Источном Сарајеву, Филозофски факултет Пале.

Ковачевић, М. (2015). *Стилистика и граматика стилских фигура*. Београд: Јасен.

Ковачевић, М. (2020). Типови говора у роману Џејн Остин „Разум и осјећајност". *Филолог–часопис за језик, књижевност и културу*, (22), 116–142.

Kurfalı, M. & Wirén, M. (2020). Zero-shot cross-lingual identification of direct speech using distant supervision. *Proceedings of LaTeCH-CLfL 2020*, 105–111.

Leech, G., & Short, M. (2007). [1981]. *Style in fiction: A linguistic introduction to English Fictional prose*. 2nd ed. Harlow, Essex: Pearson Education.

Mahlberg, M., Wiegand, V., Stockwell, P., & Hennessey, A. (2019). Speech-bundles in the 19th-century English novel. *Language and Literature*, *28*(4), 326–353. https://doi.org/10.1177/0963947019886754

Nykänen, E., & Koivisto, A. (2016). Introduction: Approaches to fictional dialogue. *International Journal of Literary Linguistics* 5(2), 1–14.

Oostdijk, N. (1990). The language of dialogue in fiction. *Literary and Linguistic Computing* 5(3), 235–241.

Panić Cerovski, N., & Ivanović, B. (2017). Glagoli govorenja kao markeri emfaze – slučaj pseudocitiranja. *Komunikacija I Kultura Online*, *7*(7), 142-154. https://doi.org/10.18485/kkonline.2016.7.7.10

Savić, S., & Polovina, V. (1989). *Razgovorni srpskohrvatski jezik.* Novi sad: Filozofski fakultet, Institut za južnoslovenske jezike.

Semino, E., Short, M., & Wynne, M. (1999). Hypothetical Words and Thoughts in Contemporary British Narratives. *Narrative*, *7*(3), 307–334.

Semino, E., & Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Polovina, V. (1988). The Basic "Verba Dicendi" and their cohesive role in spoken conversation. *Acta Linguistica Hungarica*, *38*(1/4), 193–200.

Polovina, V. (1999). *Semantika i tekstlingvistika*. Beograd: Čigoja štampa.

Polovina, V. (2021). Non-omission of Deictic Personal Pronoun, In: Polovina, V. & Panić Cerovski, N. (Eds.), *BeLiDa 2021* (pp. 19–20). Belgrade: Faculty of Philology.

Thomas, B. (2012). *Fictional Dialogue: Speech and Conversation in the Modern and Postmodern Novel*. Lincoln, NE: University of Nebraska Press.

Toolan, M. (1985). Analyzing Fictional Dialogue. *Language and Communication*, *5*(3), 193–206.

Toolan, M. (1987). Analysing Conversation in Fiction: The Christmas Dinner Scene in Joyce's: "Portrait of the Artist as a Young Man." *Poetics Today*, *8*(2), 393–416.

Toolan, M. (2001). *Narrative: a critical linguistic introduction.* (2nd ed.). London: Routledge.

Weiser, S. & Watrin, P. (2012). Extraction of unmarked quotations in Newspapers; A Study Based on Direct Speech Extraction Systems. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 559–562.

Wirén, M., Ek, A., & Kasaty, A. (2020). Annotation Guideline No. 7: Guidelines for annotation of narrative structure. *Journal of Cultural Analytics*. doi: 10.22148/001c.11772

Yeung, C. Y., & Lee, J. (2016). An annotated corpus of direct speech. In: Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Paris, France. European Language Resources Association (ELRA), 1059–1063.

Yeung, C. Y., & Lee, J. (2017). Identifying speakers and listeners of quoted speech in literary works. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Vol 2: 325–329. Asian Federation of Natural Language Processing.

***Analyzed Corpus*[15]:**

NP – Rooney, S. (2018). *Normal People*. London: Faber & Faber Ltd.

NLj – Runi, S. (2019). *Normalni ljudi*. Beograd: Geopoetika izdavaštvo.

Used Tools:

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

[15] Corpus in .txt form was constructed only for the purpose of the research.

Милица Динич Маринкович

(НЕКОТОРЫЕ ПРОСТЫЕ) СРЕДСТВА ДЛЯ ИЗВЛЕЧЕНИЯ ОРФОГРАФИЧЕСКИ НЕМАРКИРОВАННОГО ХУДОЖЕСТВЕННОГО ДИАЛОГА

Р е з ю м е

Одно из простых упражнений для студентов любого курса корпусной лингвистики состоит в том, чтобы сформировать регулярное выражение, которое будет соответствовать вымышленным строкам диалога / прямой речи в некоторым корпусам художественной прозы. Это довольно удобная задача, поскольку решение этой проблемы заключается в идентификации и определении орфографических и типографских соглашений для представления прямой речи в данном корпусе и не требует аннотации корпуса. Но вопрос в том, что происходит, когда эти орфографические и типографские маркеры в тексте отсутствуют?

Чтобы дать ответ на этот вопрос, в статье представлен краткий обзор последних подходов и предлагаемых решений этой проблемы, с особым вниманием к работе с неаннотированным корпусом. Выбранные подходы и предлагаемые решения основаны на процессах машинного обучения, различных наборах эвристик для автоматического извлечения, а также аннотации корпусов или создании электронных словарей. Тем не менее, основная цель этой исследовательской работы состоит в том, чтобы выяснить, возможно ли вернуть эту проблему в рамки простых студенческих упражнений и в какой степени.

Таким образом, простые средства извлечения, которые мы предлагаем, основаны на идентификации лингвистических особенностей строк в орфографически и типографически немаркированных художественных диалогах и различных способах их преобразовании в шаблоны для извлечения из неаннотированных корпусов. В качестве исследовательского корпуса мы использовали роман Салли Руни „Нормальные люди”, как оригинальную версию (на английском языке), так и сербский перевод.

Ключевые слова: немаркированный художественный диалог, немаркированная прямая речевая репрезентация, лингвистические особенности художественного диалога, образцы извлечения из неаннотированных корпусов, художественный диалог Салли Руни