

# Quotation and Narration in Contemporary Popular Fiction in Swedish – Stylometric Explorations

Mats Dahllöf<sup>1</sup>

<sup>1</sup>*Uppsala University, Uppsala, Sweden*

## Abstract

A fundamental feature of many genres of fiction is the alternation between a narrative frame (NF) and quoted inset (QI) dialogue. Both formal and representational features distinguish NF and QI segments. This paper is an explorative study on the stylistic differentiation between frame and inset material in recent commercially successful fiction in Swedish. There are mainly two orthographic options as regards this distinction: Explicitly enclosing inset segments within quotation marks is one. Using an initial dash to indicate utterance display is another, in which case frame and inset material typically alternate in a way not made explicit by the orthography. The corpus behind the present study comprised 450 novels. In order to deal with dash orthography data (135 books), we trained a multilayer perceptron classifier to tell NF and QI segments apart. We relied on the fact that native quotation mark text can be converted to annotated dash orthography data, which can then be used for supervised training and validation. A small-scale manual evaluation on the texts we aim to analyze, yielded an accuracy score around 95%. In order to explore the stylometric relations between NF and QI components in the novels, we looked at a selection of basic grammatical features. A characterization of each feature was made by means of recording the fraction of works in which the relative frequency of the feature is higher in QI than in NF. This summarizes how authors tend to “use” that feature to create a contrast between NF and QI. Another way to examine how the NF and QI styles are related is to apply a correlation test. We then saw, for instance, that QI material in 100% of the books are denser in auxiliary verbs, second person pronouns, and interjections, while NF segments in all or almost all cases are denser in nouns, adjectives, third person pronouns, and prepositions. We could also observe that e.g. noun density in NF and QI correlate in a strong way. The same holds for adverbs and cardinal numerals. This suggests that books and authors exhibit stylistic tendencies which affect both narrator and the characters, as far as the kind of fiction we have studied go.

## Keywords

fiction, dialogue, quotation, quoted inset, narration, narratorial frame, direct speech, stylometry

## 1. Introduction

The present study is concerned with recent commercially successful fiction in Swedish. It will focus on a stylometric exploration of the differentiation between frame narration and quoted inset material, typically dialogue. A component in this work was the training of a classifier which can tell the two kinds of content apart in novels written in “dash orthography”, which makes this a non-trivial task.

---


*The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.*

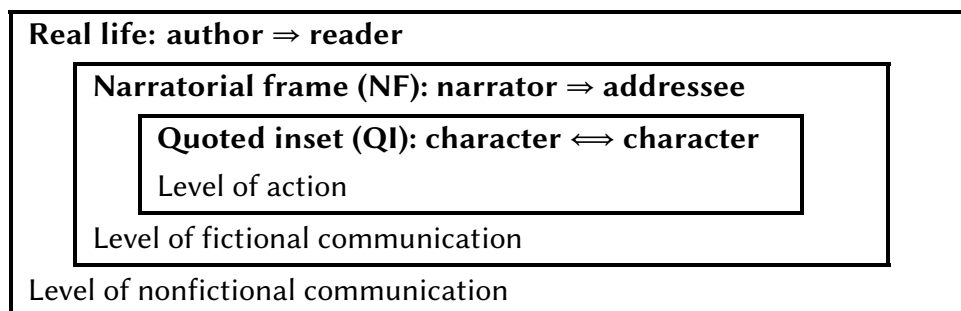
✉ [mats.dahllof@lingfil.uu.se](mailto:mats.dahllof@lingfil.uu.se) (M. Dahllöf)

🆔 0000-0002-4990-7880 (M. Dahllöf)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Frame and inset [1] in a “Chinese boxes” model [2] of the levels of communicative contact.

A fundamental feature of many genres of fiction is the alternation between a narrative frame (NF) and quoted inset (QI) dialogue, see Figure 1. Dialogue is the mimetic, more or less verbatim display – using the device of “direct speech” –, of utterances made by the characters in the story as chains of events play out. QIs are distinguished by both formal and representational features. One of the former is directness, i.e. “the inset’s syntactic and deictic independence of the frame” [1, p. 111]. The distinction between quoted dialogue and other modes of narration is not always a matter of a simple binary opposition, but the analysis here will treat it as such. This simplification is for the most part adequate for the kind of fiction we analyse here. It is also possible for authors to use indirect speech and free indirect speech – as illustrated in examples (3) and (4), respectively, below –, to mention the two most well-known alternatives. Speech acts can also be described in ways that only summarize their content or focus on other aspects (cf. e.g. [3] and [4]). Furthermore, we should remember that both narration and quotation – and indeed any kind of communication – can be embedded in QI material.

In this article we will address two research questions in an exploratory fashion. The first one is instrumental: How well can we separate inset and frame text automatically? Secondly, we will investigate whether and how a systematic exploration of stylometric features can be used to find and illustrate differences between inset and frame text as well as correlations between the two embedded styles over a corpus of novels. The method and aims behind this study belong to the school of “distant reading [...] focus[ing] on units that are much smaller or much larger than the text: devices, themes, tropes” [5].

In Swedish fiction, there are two main orthographic options for differentiating between NF and QI material. The most common and most explicit one is enclosing quoted segments within quotation marks (“”). The quotation marks serve as an explicit markup unambiguously indicating the start and end of directly quoted material. Example (1) illustrates this, with the official English translation exhibiting the explicit style. Using a dash (–) to indicate the start of an utterance display paragraph – as in the original (2) – is a less explicit and somewhat less common (30% of the books in our corpus, see below) orthographic device.

- (1) “För många träffar”, säger han. ”Tiden rinner ut.” [Adapted.]  
 “Too many results,” he says. “Time’s running out.” [Official English translation.<sup>1</sup>]

<sup>1</sup>Lars Kepler, *Stalker*, Alfred A. Knopf, 2016. Translation by Neil Smith.

- (2) – För många träffar, säger han. Tiden rinner ut. [Lars Kepler, *Stalker*, Bonniers, 2014.]  
– Too many results, he says. Time’s running out. [Adapted.]
- (3) He said that there were too many results and that time was running out. [Adapted.]
- (4) [He was pessimistic.] There were too many results. Time was running out. [Adapted.]

In dash-style novels we often find an unmarked alternation between frame and inset segments in the same paragraph, which would have been unambiguous had quotation marks been used. So, the “inquit” formula *he says* in (2) belongs to the NF, but could also – quite easily if you disregard the wider context – be understood as being part of the QI. More extended frame material surrounded by quotation is known as “suspended quotation” [6]. Identifying the spans of QI segments in dash-style fiction consequently relies on semantic and pragmatic factors and requires an interpretational effort from the reader. It also constitutes a non-trivial problem for engineering in natural language processing (NLP).

Supplementary Materials providing details on the corpus and a more complete presentation of results is available at URL <https://github.com/mdahl11of/dhnb2022sm>.

## 2. Previous Research

The analysis of dialogue and narrative goes back to Plato’s opposition between *μίμησις* and *διήγησις*, or between showing and telling, in modern critical parlance. Sternberg [1] lists five representational features associated with direct speech and with the idea of *μίμησις* as “speak[ing] in the person of another” [1, p. 111]: It is empathetic, specific, realistic, distinctive, and reproductive. Quoted speech “is conventionally understood to replicate exactly what the quoted character is supposed to have said” [4]. By contrast, two of the most salient features of narratives are their temporal relations to the events of the story and the “voice” of the narrator. Authors thus have a range of options relating to the narration, e.g. first or third person perspective and past or present tense. (See [3] for an extended discussion.)

Still, it has often been assumed in corpus-based literary studies that fiction is one register and that a novel, for instance, exhibits one style. Egbert and Mahlberg [7] reject this idea. They use multidimensional analysis involving factor analysis on a corpus of 19th-century English novels to show that there are “extreme differences between the linguistic characteristics of fictional speech and narration” [p. 86], which are two register categories typically “interspersed throughout a text” [p. 98].

Quoted inset recognition as a problem in NLP has been addressed in a number of studies, which have been motivated by the central importance of the task for computational literary studies. Recent work, e.g. [8], [9], [10], and [11], rely on machine learning but also involve rule-based procedures. A closely related issue for NLP engineering is identifying the speakers behind inset segments [12].

## 3. Data

The corpus behind the present study comprises a collection of bestsellers and beststreamers in Swedish, including translations, in the Swedish book market 2015–2020, along with a batch of

**Table 1**

Some corpus statistics. The quotation mark and dash novels ( $n = 450$ ) form the data for the present study.

	Bestseller (313)			Beststreamer (56)		Storytel Original (98)		$\Sigma$
	Crime	Prestige	Other	Crime	Other	Crime	Other	
quotation mark	110	8	77	35	2	37	46	315
dash	75	9	18	18	1	6	8	135
other	0	8	8	0	0	0	1	17
$\Sigma$	185	25	103	53	3	43	55	467

original productions in Swedish from the online streaming platform Storytel. The bestsellers are the novels which have earned that status according to the Swedish Publishers' Association (SvF), either in hardback or paperback. As beststreamers (cf. [13] and [14]) we count the novels (excluding the categories history, biographies, teens and young adult, and children's books) which have been among the twenty most streamed audiobooks each year in the Storytel platform. Storytel Original works are primarily produced for online streaming consumption ("born-audio"), but are also made available as e-books. The current corpus includes all adult fiction in Swedish published by Storytel Original from its inception in 2016 until May 2021.<sup>2</sup> The corpus is thus not a sample, but a complete dataset given the parameters defining it. As regards genre, we distinguished crime, prestige (which does not overlap with crime for any of the books we have looked at), and other genres. See [14] for details on genre assignment. Table 1 gives an overview of the composition of the corpus. A complete list is provided in the Supplementary Materials.

In the classifier training and the stylometric analysis we distinguished the following three categories of paragraph.

- NF only paragraphs, i.e. those not beginning with a dash (–) or containing quotation marks ("). Henceforth: *NF paragraphs*.
- Typical QI paragraphs, i.e. those beginning with QI material as indicated by an initial quotation mark or dash. They are often QI only, but there is often an alternation between QI and NF text. Henceforth: *QI/NF paragraphs*. (Our classifier targeted the QI/NF separation in the dash-orthography variety of such paragraphs.)
- Other paragraphs, i.e. those with non-initial explicitly quoted text or non-matching quotation marks. They were excluded from the quantitative analysis.

The machine learning design, as well as the stylometric analysis, departed from data tagged with part-of-speech (POS) labels. We used the Stanza [15] system and its "talbanken 1.0.0" [16] model for Swedish<sup>3</sup> to tag the texts.

<sup>2</sup>Many of these works are available in "seasons" comprising 10–24 episodes. The length of each season is roughly that of a novel, and we concatenated all the episodes of each season into one document in this study. The only Storytel Original among the beststreamers was a first episode of a season, *Svart stjärna – S1E1* (2016), by Jesper Ersgård and Joakim Ersgård. That text is counted as a part of a Storytel Original in the corpus.

<sup>3</sup><https://stanfordnlp.github.io/stanza/>.

## 4. Quoted Inset vs Narratorial Frame Classification

Separating QI and NF text in dash-style QI/NF paragraphs is a classification problem. Transforming text written in the quotation mark style, e.g. (1), into the dash style, e.g. (2), is for the most part trivial. This means that native quotation mark text can be converted into annotated dash-style data. We can use such minimally artificial data for supervised training and validation of classifiers working on dash style documents (as do e.g. [8] and [10]). These classifiers can then be used to separate NF and QI text in paragraphs originally written in the dash style.

We did not explore the possibilities for classifier design in depth, but managed to find an approach which can be considered satisfactory, at least for the purpose at hand.

The paragraphs were segmented at punctuation marks, where QI/NF shifts may occur. This leaves us with short snippets to be classified. Features were derived from both the snippet text and the preceding narration context, and involved POS unigrams and bigrams, word forms, as well as pronoun person and verb tense. This allows the classifier to take advantage of both lexical information and point of view differences between the narration and the dialogue. A multilayer perceptron with three layers, each comprising 15 neurons, was found to be the optimally performing architecture, with 800 features selected. (The classifier pipeline can be found among the Supplementary Materials.)

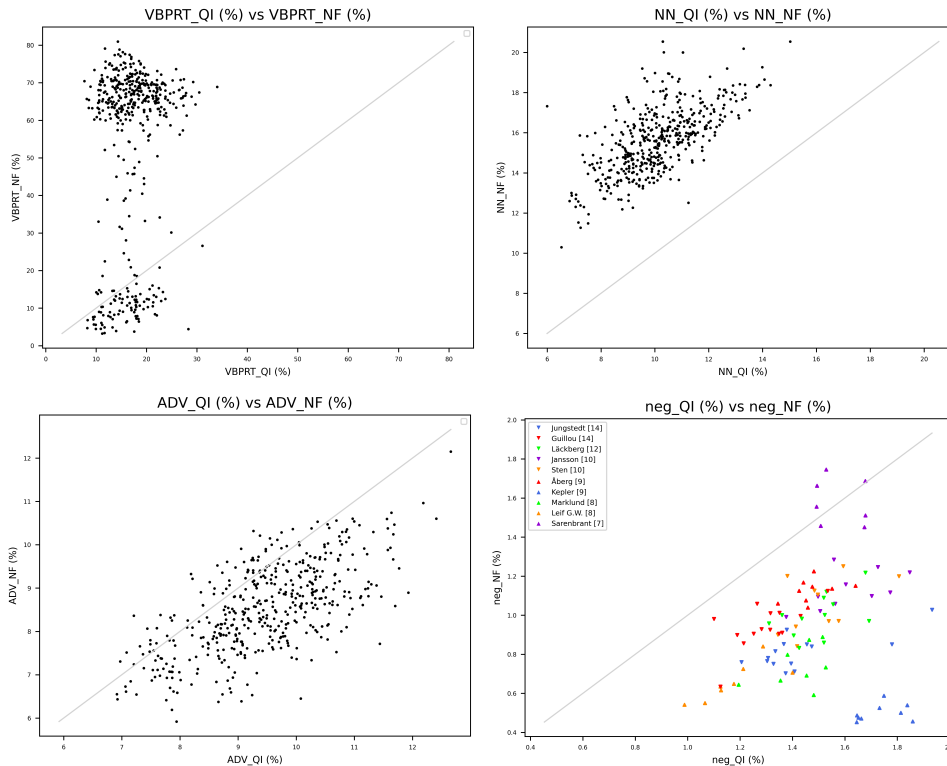
We applied a cross-validation setup which was three-fold over 315 books (see first row in Table 1) taking, for training, 80,000 randomly selected snippets from 210 books and as testing data all snippets from all dialogue paragraphs from the 105 other books. We then saw accuracy scores from 94.2% to 94.6% (on word token level: 93.2% to 93.4%).<sup>4</sup> A small-scale manual evaluation of the best-performing of the three classifiers on the kind of data we actually aim to analyze, i.e. dash-orthography novels, yielded an accuracy score of 95.7%, based on 282 randomly selected snippets from the 135 relevant novels in our corpus. This performance is difficult to compare with previous approaches, but we can note that [8] hold that “an accuracy of 0.9 is remarkable”. The resulting classifier was then used to tag the 135 books for the stylometric analysis discussed below.

## 5. Stylometric Exploration

In order to explore the stylometric relations between NF and QI components in the novels, we took a selection of very basic grammatical features (listed in Table 2) as our point of departure. They comprise parts of speech, pronouns according to grammatical person, and negation (only *inte*). These were quantified as relative frequency among word tokens. We also included the most important inflectional categories, quantified as ratio in tokens of the applicable part of speech. (The full table of novels times feature values is available in the Supplementary Materials.) A survey of a population that includes all elements defined by a set of criteria, like the present one, does not face a problem of sampling bias. This means that simple descriptive statistics are relevant, while significance testing is not applicable as in sample-based studies [17].

---

<sup>4</sup>On the level of single books, the score is between 79.5%, for *Kvinna inför rätta* (*Apple Tree Yard*) by Louise Doughty, and 99.5% for *Playground* by Lars Kepler and *Tärtgeneralen* by Filip Hammar and Fredrik Wikingsson. The Doughty novel is a first/second person singular narrative in present tense, while the two with the highest scores are in a third person past tense NF.



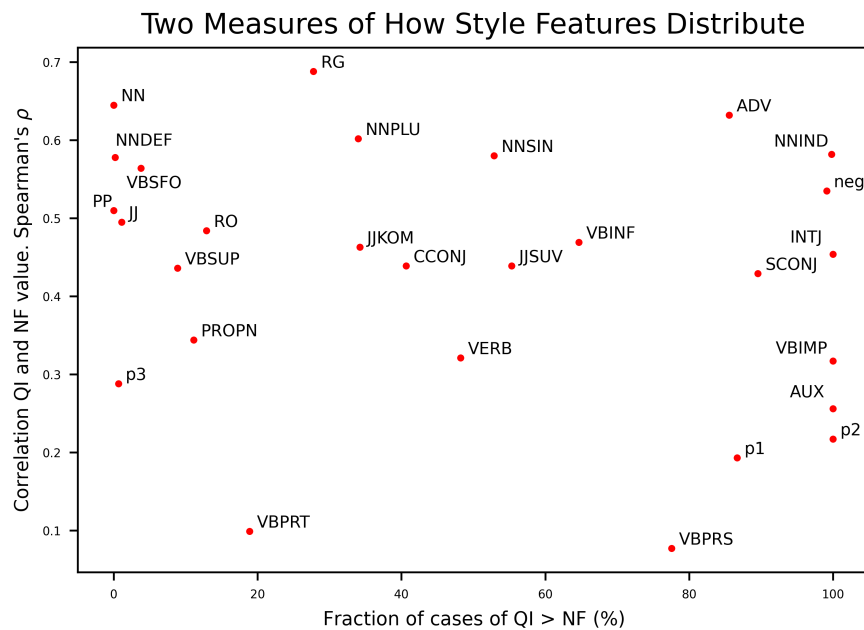
**Figure 2:** QI against NF values plotted for four features. Negation restricted to the ten Swedish writers with the largest number of works in the corpus – as an example of author-specific idiosyncrasies of style. The grey diagonal shows where the QI and NF values are equal.

A direct way of visualizing the QI and NF values for the various books and features is to scatter-plot them against each other as in Figure 2, which shows four examples of features. The NF values for preterite clearly distribute the works bimodally, forming a larger high-value cluster and a smaller low-value cluster. The QI values are clearly less dispersed, suggesting that authors to a high extent “agree” on the proper average amount of preterite in fictional speech. This does not correlate to any visible degree with the preterite density in the NF. (The plot for present tense, see Supplementary Materials, presents a similar shape upside down.) Looking at another plot, we see that the noun density is quite dispersed both in NF and QI. However, the NF value is always higher than the QI value, even if the ranges of possible values overlap. Furthermore, the QI and NF values correlate. In other words, there is a tendency for noun-dense narrators to be coupled with noun-dense characters. We see the same kind of correlation for adverbs, but in most books (86% of them) the QI material is denser in adverbs than the NF. Finally, in Figure 2, there is a plot for negation only showing books by the ten most prolific Swedish authors in the corpus. The distribution is roughly as for adverbs in general, but we also see that most authors cluster in particular regions, i.e. they use negation consistently in their books in both NF and QI. Lars Kepler, for instance, is low-negation in NF, but high in QI. Sofie Sarenbrant is high-negation in both NF and QI, even extremely so in NF.

**Table 2**

Fractions of works in which the QI value exceeds the NF value for the features explored and Spearman's  $\rho$  for their correlation.

Feature	QI > NF (%)	$\rho$	Feature	QI > NF (%)	$\rho$
VERB (full lexical verb)	48	0.32	p1 (first person pronoun)	87	0.19
AUX(iliary verb)	100	0.26	p2 (second person pronoun)	100	0.22
NN (noun)	0	0.64	p3 (third person pronoun)	1	0.29
PROPN (proper noun)	11	0.34	neg(ation)	99	0.53
JJ (adjective)	1	0.49	VBPRT (preterite)	19	0.10
RG (cardinal numeral)	28	0.69	VBPRS (present)	78	0.08
RO (ordinal numeral)	13	0.48	VBSUP (supine)	9	0.44
PP (preposition)	0	0.51	VBIMP (imperative)	100	0.32
ADV(erb)	86	0.63	VBINF (infinitive)	65	0.47
CCONJ (coordinating conjunction)	41	0.44	VBSFO (s-form)	4	0.56
SCONJ (subordinating conjunction)	90	0.43	NNIND (indefinite)	100	0.58
INTJ (interjection)	100	0.45	NNDEF (definite)	0	0.58
JJSUV (superlative)	55	0.44	NNSIN (singular)	53	0.58
JJKOM (comparative)	34	0.46	NNPLU (plural)	34	0.60

**Figure 3:** The values in Table 2 plotted.

A straightforward characterization of the “behaviour” of each feature was made by means of recording the fraction of works in which the relative frequency of the feature is higher in QI than in NF (i.e. above the diagonals in Figure 2). This score summarizes how authors tend to “use” that feature to create a contrast between NF and QI material. Another way to examine how the NF and QI styles are related is to apply a correlation test. Not wishing to assume that

these variables are normally distributed, we used Spearman's correlation coefficient ( $\rho$ ), which is rank-based and consequently non-parametric (i.e. not based on a normality assumption). Noun and adverb densities, as shown in Figure 2, will yield quite high  $\rho$  values.

Table 2, whose values are plotted in Figure 3, shows the results. We find, as it were, a left column of NF-oriented features and a right one of QI-oriented ones. In eight cases the tendencies hold for all novels (0% or 100%) in the corpus. The NF-oriented features comprise elements of general and definite nominal reference, prepositions, passive and past tense verb forms, and numerals. The QI material, by contrast, is generally denser in interjections, first and second person pronouns, and present tense and imperative verb forms. Nouns are indefinite to a higher degree than in the NF. A plausible explanation is that there is less of anaphoric reference to already introduced entities. This is as can be expected from dialogue. The higher incidences of auxiliary verbs, adverbs, including negation, and subordinating conjunctions does not as immediately chime with what can be expected.

We see that there is a positive correlation between QI and NF values for all features, but the tendency is of a stronger kind ( $\rho > 0.55$ ) as regards nouns, cardinal numerals, adverbs, and the inflectional categories of nouns. The lowest degree of correlation is seen in features related to deictic reference, i.e. tense and grammatical person, whose use in narration or dialogue is based on pragmatic principles.

## 6. Discussion and Conclusions

We have introduced a method for visualizing and quantifying how authors use various grammatical features to differentiate between NF and QI material in fiction. Our observations clearly agree with the findings of Egbert and Mahlberg [7], according to whom present tense and first and second person are associated with dialogue, while there is more of past tense and third person in narration. These results are far from unexpected and confirm the sanity of their and our methods. Our analysis clearly revealed tendencies which are strong, but also of a fairly abstract kind. There is surely much more to find in the data.

We have found that NF and QI segments can be told apart with a high degree of accuracy by means of a multilayer perceptron. It should be stressed that there is a potential methodological danger in using a classifier that to a large extent rely on the same kind of features that are explored in the study. As always, patterns revealed by corpus-based "distant reading" should be seen as an invitation to look closer at the data.

Another open and interesting issue is to what extent the choice of QI orthography influences other stylistic options. The absence of quotation marks in the dash orthography is likely to prompt authors to use other means of making the QI status of segments obvious to the readers.

The present study has, in a sense, provided a map to the comfort zones of Swedish popular fiction and to the vast and fascinating regions of stylistic space this genre tends to shun.

## Acknowledgments

This work was carried out in the project "Patterns of Popularity: Towards a Holistic Understanding of Contemporary Bestselling Fiction" funded by Vetenskapsrådet (2019-02829). PI Karl



Berglund contributed to the design and data curation for the present study.

## References

- [1] M. Sternberg, Proteus in quotation-land: Mimesis and the forms of reported discourse, *Poetics Today* 3 (1982) 107–156.
- [2] M. Jahn, *Narratology 2.3: A guide to the theory of narrative*, 2021. URL: <https://www.uni-koeln.de/~ame02/pppn.pdf>.
- [3] S. Rimmon-Kenan, *Narrative Fiction: Contemporary Poetics*, Routledge, London and New York, 1983.
- [4] B. McHale, Speech representation, in: P. Hühn, J. Pier, W. Schmid, J. Schönert (Eds.), *the living handbook of narratology*, Hamburg University, Hamburg, 2014.
- [5] F. Moretti, Conjectures on world literature, *New Left Review* 1 (2000) 54–69.
- [6] J. B. Herrmann, A. M. Jacobs, A. Piper, Computational stylistics, in: D. Kuiken, A. M. Jacobs (Eds.), *Handbook of Empirical Literary Studies*, De Gruyter Reference, 2021.
- [7] J. Egbert, M. Mahlberg, Fiction – one register or two? speech and narration in novels, *Register Studies* 2 (2020) 72–101.
- [8] F. Jannidis, L. Konle, A. Zehe, A. Hotho, M. Krug, Analysing direct speech in german novels, in: 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHD 2018, Köln, Germany, February 26 - March 2, 2018, 2018.
- [9] A. Ek, M. Wirén, Distinguishing narration and speech in prose fiction dialogues, in: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, 2010, pp. 124–132.
- [10] M. Kurfalı, M. Wirén, Zero-shot cross-lingual identification of direct speech using distant supervision, in: *Proceedings of LaTeCH-CLfL*, 2020, pp. 105–111.
- [11] J. Byszuk, M. Woźniak, M. Kestemont, A. Leśniak<sup>1</sup>, W. Łukasik, A. Šeĵa, M. Eder, Detecting direct speech in multilingual collection of 19th-century novels, in: *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, 2020, pp. 100–104.
- [12] D. K. Elson, K. R. McKeown, Automatic attribution of quoted speech in literary narrative, in: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, 2010, pp. 1013–1019.
- [13] K. Berglund, Introducing the beststreamer: Mapping nuances in digital book consumption at scale, *Publishing Research Quarterly* 37 (2021) 135–151.
- [14] K. Berglund, M. Dahllöf, Audiobook stylistics: Comparing print and audio in the bestselling segment, *Journal of Cultural Analytics* 11 (2021) 1–30. doi:10.22148/001c.29802.
- [15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1001–1008.
- [16] J. Nivre, A. Smith, UD Swedish Talbanken, 2008. URL: [https://universaldependencies.org/treebanks/sv\\_talbanken/index.html](https://universaldependencies.org/treebanks/sv_talbanken/index.html).
- [17] N. Hirschauer, S. Grüner, O. Mußhoff, C. Becker, A. Jantsch, Can p-values be meaningfully interpreted without random sampling?, *Statistics Surveys* 14 (2020) 71–91.