Artjoms Šeļa, Ben Nagy, Joanna Byszuk,
Laura Hernández-Lorenzo, Botond Szemes, and Maciej Eder

From Stage to Page: Stylistic Variation in Fictional Speech

Abstract: Stylometry is mostly applied to *authorial* style. More recently, researchers have begun investigating the style of *characters*, finding that although there is detectable stylistic variation, the variation remains within authorial bounds. In this article, we address the stylistic distinctiveness of characters in drama. Our primary contribution is methodological; we introduce and evaluate two non-parametric methods to produce a summary statistic for character distinctiveness that can be usefully applied and compared across languages and times. This is a significant advance – previous approaches have either been based on pairwise similarities (which cannot be easily compared) or indirect methods that attempt to infer distinctiveness using classification accuracy. Our first method is based on bootstrap distances between 3-gram probability distributions, the second (reminiscent of 'unmasking' techniques) on word keyness curves. Both methods are validated and explored by applying them to a reasonably large corpus (a subset of DraCor): we analyze 3301 characters drawn from 2324 works, covering five centuries and four languages (French, German, Russian, and the works of Shakespeare). Both methods appear useful; the 3-gram method is statistically more powerful, but the word keyness method offers rich interpretability. Both methods are able to capture phonological differences such as accent or dialect, as well as broad differences in topic and lexical richness. Based on exploratory analysis, we find that smaller characters tend to be more distinctive and that women are cross-linguistically more distinctive than men, with this latter finding carefully interrogated using multiple regression. This greater distinctiveness stems from a historical tendency for female characters to be restricted to an 'internal narrative domain' covering mainly direct discourse and family/romantic themes. It is hoped that direct, comparable statistical measures will form a basis for more sophisticated future studies, and advances in theory.

Artjoms Šeļa, Ben Nagy, Joanna Byszuk, Maciej Eder, Polish Academy of Sciences
Laura Hernández-Lorenzo, University of Seville
Botond Szemes, Institute for Literary Studies Budapest
Artjoms Šeļa, University of Tartu

https://doi.org/10.1515/9783111071824-008

1 Introduction

Since Vladimir Propp's work, structural narratology has approached fictional characters mainly through their role or function – by what they do or what is done to them (Eder et al. 2010). This character typology relied on recurring functions in the narrative (lover, villain, victim, detective, etc.) and the same perspective was often adopted in computational research, where characters in novels were modeled on the basis of narrative passages rather than dialogue (Bamman et al. 2014; Bonch-Osmolovskaya and Skorinkin 2017; Underwood et al. 2018; Stammbach et al. 2022).

In dramatic texts, however, the dominant device for characterization is an utterance. While the script usually contains some stage directions, the specifics of characterization and style of performance are not determined by the text itself, but developed by a specific theater, director or a troupe. Over the course of history, many plays were written for specific theater stages, and it was common practice to write characters for specific actors (Fischer-Lichte 2002). Of course, this kind of 'outsourced characterization' was supported by dramatic conventions and formulas. Viewers' expectations could be shaped without a single word being uttered on stage, just by a character wearing a costume, operating a puppet, or changing a dell'arte stock mask. At the same time, the things characters say and how they say them are the main textual source of information about them. It is reasonable to assume that dramatists make significant efforts to create linguistic distinctions between princes and paupers, lovers and schemers, aristocrats and merchants. Tragic monologue is written differently from a comedic exchange between servants. Some previous computational works treat linguistic distinctiveness of characters from the perspective of this stylistic continuum (Vishnubhotla et al. 2019), noting that it can be influenced by genre, character gender, or their social and professional dispositions.

A parallel narratological tradition, tied to Bakhtin's ideas of heteroglossia, focuses not on abstract character roles, but on the words characters say (Sternberg 1982; Culpeper 2001; Bronwen 2012). The modern novelistic space of dialogic exchange, 'educated conversation' (Moretti 2013, p. 20) and the clash of styles in reported discourse become central here. Available stylometric research on fictional speech and micro-stylistic variation suggests that characters within a text are often distinguishable by their local linguistic patterns without obscuring the global authorial trace (Burrows 1987; Hoover 2017). As put by Burrows and Craig: "Characters speak in measurably different ways, but the authorial contrasts transcend this differentiation. The diversity of styles within an author always remains within bounds" (Burrows and Craig 2012, pp. 307–308).

Conceptually and methodologically, the majority of previous works examined not the *distinctiveness* of characters, but their (pairwise) *similarity*. Similarity measures are meaningful in pairwise contexts but cannot be analyzed and compared as individual summary statistics. Since Burrows' seminal study of speech patterns in Jane Austen's characters (Burrows 1987), these approaches focused on calculating similarity within a collection of characters: how different is character X from character Y, and each of them from character Z. Burrows measured the correlation between characters' usage of 30 most frequent words (technically, he fit a linear regression for two sets of log-frequencies); later, similarity was most often inferred through clustering based on pairwise distance calculations (Reeve 2015; Craig and Greatley-Hirsch 2017; Hoover 2017). Sometimes linguistic similarity served as a basis for arguing functional similarity as well. A recent study that linked Bakhtin's dialogism and the stylistic diversity of characters' speech (Vishnubhotla et al. 2019) proposed the analysis of distinctiveness rather than similarity using supervised classification. Instead of using a network of pairwise relationships, the authors asked how well a classifier can recognize character X as being written by author A. Classification accuracy in this scenario becomes an explicit summary statistic for distinctiveness that can be assigned to a character (or, in an aggregated manner, to a play or an author). However, the supervised approach, proposed by Vishnubhotla et al., is data hungry: it suffers from extreme class imbalance, an abundance of short samples (most characters speak only a little), and is dependent on language-specific feature construction procedures.

By contrast, this paper will present a simple, non-parametric measure of character distinctiveness that is based on bootstrapped probability distributions representing a character and all others present in a given play: an approach largely informed by authorship verification techniques. This measure is language-independent and relies only on the context of a single work, which, in turn, minimizes problems of language variation, authorial signal and chronological change in a comparative setting. Individual distinctiveness scores can then be tested against other measures and metadata categories in a hypothesis-driven manner, not only across languages, but also across genres (e.g., novel vs. drama). Do comedies tend to employ more distinct characters? Does distinctiveness increase (authors get better), or decrease (social and linguistic homogenization occurs) over time? Is there a difference between the distinctiveness of fictional women and men? If so is it the direct result of perceived gender differences, or is it constructed by imagined differences in social and professional status?

Lacking good descriptive metadata on the dramatic characters, this paper will not answer the above-mentioned questions in any satisfying way. Instead, we focus on presenting and justifying the measure of distinctiveness and exploring sev-

Tab. 1: A summary of the corpus. All word and 3-gram counts are for the filtered corpus (characters that speak at least 2000 words) only.

Corpus	Total Characters	Characters Analyzed	Unique 3-grams	Unique Words	Total 3-grams	Total Words
French	15462	1744	9896	79994	29.79 m	5.47 m
German	14010	1182	14341	150956	24.80 m	4.31 m
Russian	3707	248	12542	71217	4.05 m	0.72 m
Shakespeare	1431	127	5921	19595	2.16 m	0.43 m

eral factors that might shape the final scores (like the year of composition, character gender and characters' sample size).

2 Materials

As the beginning of our exploration of cross-linguistic variation, we examined four dramatic corpora from DraCor (Fischer et al. 2019): Shakespeare, French, German, and Russian. DraCor is a project that gathers dramatic corpora in various languages, primarily European, encoded in TEI-XML. With 15 corpora available so far, including the Shakespeare corpus available both in English and German, DraCor facilitates large scale analysis of dramatic conventions across language traditions and offers a wide variety of useful metadata at the level of both plays and characters. While the analysis of all DraCor corpora would be possible with the methods we developed, for the purpose of this preliminary study we focused on the languages and dramatic traditions well-known to the members of our team, eventually selecting the full corpora for Shakespeare, French, German, and Russian: a total of 2324 texts, the majority of which come from French and German. The corpus is summarized in Table 1.

3 Methods

3.1 General Approach and Definitions

Our understanding of character distinctiveness is largely informed by 'authorship verification' approaches, which center around verifying that a text is written by a target author. This problem is more general than 'authorship attribution,' which tries to identify the nearest stylistic neighbor for a text (Halvani et al. 2019). In-

stead, authorship verification asks about the relative *magnitude* of similarity: is a target text more similar to same-author samples or different-author samples? With this in mind, we define a character's 'distinctiveness' as the degree to which the style of their speech differs from that of other characters. We understand 'style' here instrumentally, as a deviation from an unobserved average language (Hermann et al. 2015), and do not introduce aggressive feature filtering, allowing both 'grammatical' and 'thematic' signals to contribute to the final measures. We anchor our distinctiveness measure in the context of the specific text in which a character appears. In theory, the frame of reference could be all plays from one author, or all plays from the same period, or even some external corpus – however, all of these would greatly complicate any comparative study.

3.2 Bootstrap 3-Gram Distinctiveness

Based on our definition of distinctiveness above, we considered a character's style to be an idiolect sampled from a frequency distribution of character 3-grams. As a natural language distribution, this was expected to be generally Zipfian, a family of heavy-tailed distributions, so non-parametric methods were seen to be important. 3-grams were preferred to words for a number of reasons: first, they capture sub-word information, which means they will reflect general sonic preferences (so they can capture things like accent) and, particularly in inflected languages, also reflect some grammatical style; second, as a practical matter, they effectively expand the sample data, since a string of text produces approximately one 3-gram per character. This increased sample size should reduce the variance of the statistics. Finally, the number of unique 3-grams in a language is considerably smaller than the number of words, so the frequency data is less sparse, which again is expected to increase robustness. To now operationalize the distinctiveness, as defined, we used standard bootstrap methods to measure the median energy distance (Székely and Rizzo 2013) with bootstrap confidence intervals between the two distributions (character 3-gram frequencies vs. 'other' 3-gram frequencies). The energy distance is one of a family of related metrics that are commonly used to measure the difference between probability distributions.

Some limitations and choices were required. As mentioned, we measured distinctiveness only within the context of a single work (even for authors with multiple works). To expand beyond single works would produce very mismatched sample sizes, since some authors were prolific and some produced just one play; even with non-parametric methods, hugely mismatched sample sizes are problematic. Furthermore, the plays span four languages and roughly five centuries, making the 'distant' context seem ridiculous. As well as the selected distinctiveness statis-

tic (median energy distance) we also recorded a 'baseline' distinctiveness, this being each character's distance from themselves. The theoretical baseline is, of course, zero, but the sample baselines will not be, meaning that this gives us an idea of the inherent variance of the samples. Finally, when selecting characters to examine, we chose a minimum size of 2000 words. Sample sizes are somewhat arbitrary, and are matters of debate (Eder 2015, 2017), but this seemed to be a reasonable, or perhaps even slightly aggressive, lower bound.

3.3 Area under Keywords

Our second, supplementary approach was informed by 'unmasking' techniques often employed in stylometric research (Koppel and Schler 2004; Kestemont et al. 2016; Plecháč and Šeļa 2021). Unmasking refers to a range of methods that share one goal: to measure and compare the *depth* of the differences between two sets of texts. For example, an author might write both high fantasy fiction and historical novels: a classifier would have little difficulty distinguishing one genre from another by simply using superficial features (e.g., 'dragons,' 'magic,' 'elves'). However, by assumption, if these most distinctive features are removed, the classifier will have more trouble determining which text came from which pool, because the texts share one deep similarity – a common authorial style. Conversely, if we compare books by two different fiction writers, these texts will also have superficial differences. However, while removing more and more distinctive features, the classifier should remain confident in distinguishing the authors from each other, because the texts do not share an authorial style that is deeply rooted in common linguistic elements and distributed over many features. By comparing the speed with which the rates of accuracy decay, we can approach authorship verification problems, i.e., how plausible is it that this text belongs to author A?

We applied the same thinking to fictional characters, as opposed to authors: the distinctiveness of a character may rely on a small number of catch-phrases ('Gadzooks!' or 'Cowabunga!') or it may be driven by non-stylistic, referential factors (Mary, speaking to John, is not likely to use the word 'Mary,' but is likely to use the word 'John,' and vice-versa). On the other hand, there are characters whose speech systematically differs from the neutral language: such as when the author imitates dialects, slang, regionalism, speech and phonetic idiosyncrasies. In the former case, an imaginary classifier should quickly lose accuracy (since John and Mary speak quite similarly), but in the latter case the removal of a small number of features would not be enough to disrupt classification.

In our case, it was impractical to use 'standard,' supervised (i.e., classifier-based) unmasking because individual characters, as samples, were simply too

small. Instead we used word keyness – a character's relative preference for a word in the context of a given drama – to calculate an alternative distinctiveness score together with a bag of easily interpretable features per character. First, we use weighted log-odds (Monroe et al. 2008) to calculate keywords for a character relative to the speech pool of the rest of the cast; second, we represented each character by their 100 words with the highest keyness, arranged by rank; finally, we measured the area under this curve, which we interpret as distinctiveness – characters with just a few key words will exhibit less area under the keyness curve. By comparing these final areas, we can compare the relative difference between each character and rest of the speech in the play. In a similar manner to the bootstrapped approach, we upsample each character's word pool to match the size of the rest of the words in the play to minimize the effect of the sample size as much as possible.

4 Results

Overall, the distinctiveness energy statistic appears useful. The baseline (character vs. self) is quite stable cross-linguistically, although it is slightly higher for characters with a very large share of dialogue (Figure 1). Note also that the distinctiveness statistic appears roughly Gaussian (see Appendix B for more discussion) and its range is relatively consistent between languages (peaking at roughly 0.20), although this consistency does not apply at the level of authors. The obvious issue is that there is a strong negative correlation between character size and distinctiveness, but this is not only a limitation of the method – lead characters naturally set the dominant style of a text (and possibly inherit more of the 'true' authorial voice). Importantly, distinctiveness does not increase with the number of speakers in a play. The method works best when there are reasonable sample sizes for both the examined character and the 'other' class. This is illustrated by the 'U' curve visible in the French corpus in Figure 1 as the examined characters' dialogue share passes 50%. As hoped, the energy-distance method does appear to capture characters who are written with distinctive idiolects, representing things like foreign accents or social class. For a discussion of this see Section 5.

As seen in Figure 2, there is no clear correlation between the date of composition and character distinctiveness, which suggests that language change does not disturb the measure. The finding that seems clear is that women are written differently from men. Female characters are generally more distinctive in all corpora (Figure 3b), although this is not visible using the keyness AUC measure – leading us to conclude that the keyness measure has lower power. This difference in the

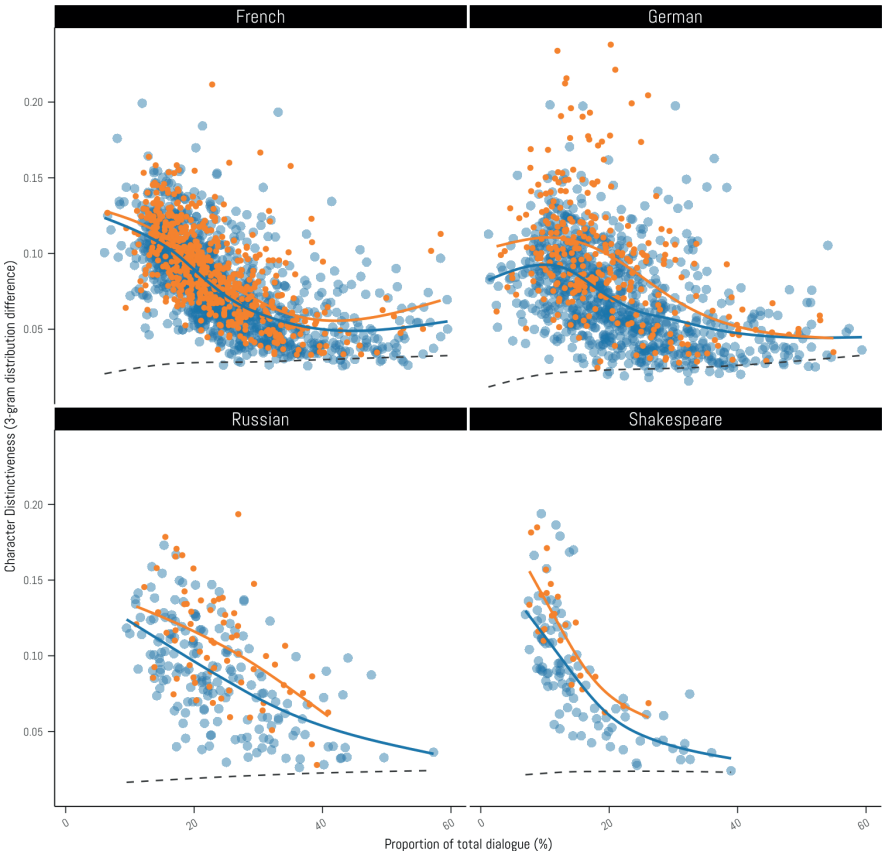


Fig. 1: Character distinctiveness, per corpus, versus % Dialogue. Women are shown smaller and in orange, men (and undefined) larger and in blue. GAM (Generalized Additive Model) trendlines are superimposed in the same colors. Baseline data (GAM trend for distinctiveness of character vs. self) is shown as a dashed line.

distinctiveness of female characters can partly be explained by the fact that they tend to have smaller parts (Fig 3a), and smaller characters in general are more distinctive (Figure 1), but that is not the whole story. Female parts have more restricted 3-gram vocabularies (Figure 3c), suggesting that they are also restricted in their semantic fields. This becomes clearer when the relative frequencies of their (word) vocabularies are examined. As well as the stereotypical tendencies (women say 'love,' men say 'sword'), the female characters, cross-linguistically, seem to be less likely to reference the 'external world' of the drama. As seen in Appendix A, relatively more frequent words for women are dominated by personal

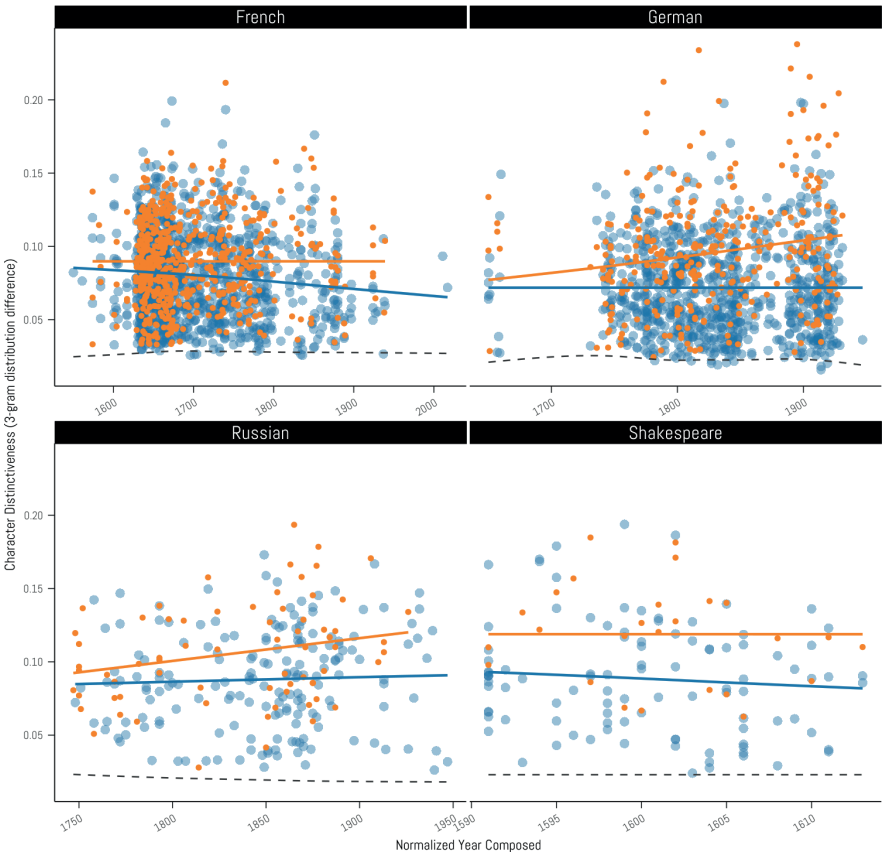
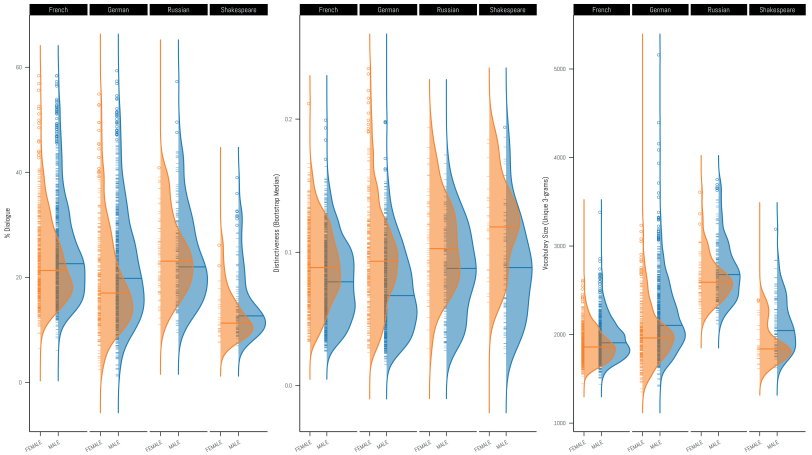


Fig. 2: Character distinctiveness, per corpus, versus year composed (DraCor data). Women are shown smaller and in orange, men (and undefined) larger and in blue. GAM (Generalized Additive Model) trendlines are superimposed in the same colors. Baseline data (GAM trend for distinctiveness of character vs self) is shown as a dashed line.

pronouns representing 'I,' 'me,' 'you' etc. or words relating to family. The male lists are dominated by indicative articles and political terms ('law,' 'noble,' 'king' etc.).

The higher distinctiveness of female characters is further supported by a formal linear model: we fit a Bayesian multiple regression where distinctiveness was conditioned on both gender and size (characters' percentage of total dialogue). A direct gender effect is present in all corpora, as expected from Figure 3a, but when we account for variation among authors, the effect may be less pronounced than it appears (for analysis and more detailed discussed, the posterior estimates are described in Appendix B). Our finding interlocks with the observation by Under-



(a) % Dialogue **(b)** Distinctiveness **(c)** Vocab size (3-grams)

Fig. 3: An analysis, per corpus, of the distribution of various features by gender. Distributions are estimated, with the median shown as a solid line. Actual points are shown as rug plots with outliers 'o' plotted for points outside 3Q + 2×IQR.

wood et al. (2018) that female characters found in English eighteenth- to twentieth-century fiction displayed high distinctiveness due to the particular way they were narrated, suggesting a pervasive authorial mentality.

5 Discussion

The measures of stylistic character distinctiveness that were proposed in this paper appear to be effective in capturing a *degree* to which characters stand out from others. The most distinctive characters, by both of our metrics, often have systematically different speech, in the form of dialects, regionalisms, or class markers. For example, Shakespeare's Captain Fluellen (*Henry V*) is Welsh and his accent is written for comedic effect. The systematic replacements b→p and d→t make him the most distinctive Shakespearean character according to both the 3-gram and word measures:

> FLUELLEN
> Your grandfather of famous memory, an't please your majesty, and your great-uncle Edward the Plack Prince of Wales, as I have read in the chronicles, fought a most prave pattle here in France.

King Henry V
They did, Fluellen.
Fluellen
Your majesty says very true: if your majesties is remembered of it, the Welshmen did good service in a garden where leeks did grow, wearing leeks in their Monmouth caps; which, your majesty know, to this hour is an honourable badge of the service; and I do believe your majesty takes no scorn to wear the leek upon Saint Tavy's day.

Regional differences also contribute to high distinctiveness in the German corpus. For example, Emerike in *das Manuskript*, written by Johanna von Weißenthurn, uses -ey instead of -ei (zwey, bey, Freylich) which is a form indicative of pre-standardized Southern German spelling. John, in Hauptmann's *Die Ratten*, speaks Plattdeutsch, a variant heavily influenced by Dutch, e.g., 'Det hat er jesacht, det ick noch ma hin müßte und janz jenau anjeben.'

In the French corpus, the most distinctive character by keyness is Gareau, from *Le Pédant Joué* (Cyrano de Bergerac), who speaks a 'patois' or rural dialect. In his critical edition, Frédéric Lachèvre comments on this distinct idiolect when Gareau is first introduced (Cyrano de Bergerac 1921, p. 25):

> Cyrano a fabriqué de toutes pièces le patois de Gareau. Le manuscrit de la BN donne un langage tout différent que celui imprimé en 1654, la pronociation des mots n'est pas tout à fait la même. Nous avons naturellement maintenu pour Gareau le texte de 1654 publié par Cyrano lui-même.
>
> Cyrano created the patois of Gareau from scratch. The manuscript of the [Bibliothèque Nationale] offers quite a different language to the one printed in 1654, the pronunciation of the words is not quite the same. We have naturally maintained for Gareau the text of 1654 published by Cyrano himself.

The most distinctive Russian characters come from Ostrovskii, who gave the main stage to Muscovite merchants and their families with their vernacular, non-aristocratic language. Tolstoy's Nikita (high on both the 3-gram and keyness lists) from *The Power of Darkness* has heavily stylized speech suggestive of Western or Southern Russian dialects, e.g., featuring a word-initial [w].

It must be borne in mind, however, that dialects or accents do not automatically cause high distinctiveness – what is being detected is the *difference* in speech patterns. In a text where everyone speaks Welsh, an English character would score highly on distinctiveness, and vice versa. Cross-linguistic inference must also account for systematic language differences: the lexical and morphological features of the various languages lead naturally to different probability distributions for

both words and N-grams (although the exact nature of those differences is too complex to grapple with here). Word-based distinctiveness measures permit easier interpretation but appear less (statistically) powerful. In addition, word-based measures operate in much higher dimensions, with all the usual problems that entails (sparsity, the 'curse of dimensionality,' etc. See, for example Moisl 2011). Finally, word-based measures naturally invite lemmatization for highly inflected languages (like Russian and German), which might cause problems for future work dealing with languages that are non-standard, historical, or otherwise less well-resourced.

We have noted that our distinctiveness measure has a strong negative correlation to the size of the character. This relationship should not be understood as a simple artifact that renders our measurement useless. Distinctive speech is always a construct, a subset of linguistic and stylistic reality. If a minor character has just a few lines about gallows and graves – like Shakespeare's gravedigger – we will never know more about their language. However, *Hamlet* is not *only* about gallows and graves; if we imagine bootstrapping the gravedigger's speech, it would be endlessly populated by these few words: we don't know how the gravedigger would speak when ruling a country or murdering their uncle. From this perspective, a protagonist is more likely to represent the lexical and stylistic norm, while minor characters will sample the Other in their ethnic, dialectal, or professional distinctiveness.

Despite the few limitations, we hope that these measures of character distinctiveness will support improved theories about style, characterization, and history. The most important question to be asked concerns the source(s) of this representational distinctiveness that authors instill in their characters. To even begin to address this issue, we need much richer annotation for characters: their social class, profession, region of origins, and age. Determining the drivers of distinctiveness will not be easy. Even to carefully verify the effect of character gender was quite complicated. We know that part of the effect comes from size: women are more likely to be minor characters. However, it is reasonable to assume that gender difference can also be confounded by genre (e.g., in comedies, there are more women playing larger roles) and social class (rural people speak more in comedies). There is also the effect of time: changing the relative dynamics of character sizes (Algee-Hewitt 2017), improving the representation of women as dramatists and altering the depiction of social class, all of which complicates the analysis even further. However, having a clear summary measure for a character's stylistic distinctiveness may help us to refine our theories on the speech of fictional characters, leading in turn to better causal models.

Funding: AŠ, JB, LHL and ME were funded by the "Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics" project (SONATA-BIS 2017/26/E/HS2/01019). BN is also grateful to QuaDramA, funded by Volkswagen Foundation and the DFG-priority programme Computational Literary Studies, for financing the presentation of the paper at the workshop.

Data and Code: The details of our approach, including data acquisition and preprocessing, were published in a Zenodo repository, allowing for full replication of all reported results: https://doi.org/10.5281/zenodo.7383687.

References

Algee-Hewitt, Mark (2017). "Distributed Character: Quantitative Models of the English Stage, 1550–1900." In: *New Literary History* 48.4, pp. 751–782. DOI: 10.1353/nlh.2017.0038.

Bamman, David, Ted Underwood, and Noah A. Smith (2014). "A Bayesian Mixed Effects Model of Literary Character." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 370–379. DOI: 10.3115/v1/P14-1035.

Bonch-Osmolovskaya, Anastasia and Daniil Skorinkin (2017). "Text Mining war and Peace: Automatic Extraction of Character Traits from Literary Pieces." In: *Digital Scholarship in the Humanities* 32.suppl_1, pp. i17–i24. DOI: 10.1093/llc/fqw052.

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

Burrows, John (1987). *Computation Into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press.

Burrows, John and Hugh Craig (2012). "Authors and Characters." In: *English Studies* 93.3, pp. 292–309. DOI: 10.1080/0013838X.2012.668786.

Craig, Hugh and Brett Greatley-Hirsch (2017). *Style, Computers, and Early Modern Drama: Beyond Authorship*. Cambridge University Press. DOI: 10.1017/9781108120456.

Culpeper, Jonathan (2001). *Language and Characterisation: People in Plays and Other Texts*. Longman.

Cyrano de Bergerac, Savinien (1921). *Les oeuvres libertines*. Ed. by Frédéric Lachèvre. Edouard Champion.

Eder, Jens, Fotis Jannidis, and Ralf Schneider, eds. (2010). *Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media*. De Gruyter. DOI: 10.1515/9783110232424.

Eder, Maciej (2015). "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." In: *Lit. Linguist. Computing* 30.2, pp. 167–182. DOI: 10.1093/llc/fqt066.

Eder, Maciej (2017). "Short Samples in Authorship Attribution: A New Approach." In: *Digital Humanities 2017. Conference Abstracts*. URL: https://dh2017.adho.org/abstracts/341/341.pdf.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke (2019). "Programmable Corpora: Introducing DraCor, an Infrastructure for

the Research on European Drama." In: *Proceedings of DH2019: "Complexities", Utrecht, July 9–12, 2019*. DOI: 10.5281/zenodo.4284002.

Fischer-Lichte, Erika (2002). *History of European Drama and Theatre*. Routledge. URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315872193-37/erika-fischer-lichte-history-european-drama-theatre-hans-van-maanen.

Halvani, Oren, Christian Winter, and Lukas Graner (2019). "Assessing the Applicability of Authorship Verification Methods." In: *arXiv:1906.10551 [cs, stat]*. DOI: 10.1145/3339252.3340508.

Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch (2015). "Revisiting Style, a Key Concept in Literary Studies." In: *Journal of Literary Theory* 9.1, pp. 25–52. DOI: 10.1515/jlt-2015-0003.

Hoover, David L (2017). "The Microanalysis of Style Variation." In: *Digital Scholarship in the Humanities* 32.suppl_2, pp. ii17–ii30. DOI: 10.1093/llc/fqx022.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans (2016). "Authenticating the writings of Julius Caesar." In: *Expert Systems with Applications* 63, pp. 86–96. DOI: https://doi.org/10.1016/j.eswa.2016.06.029.

Koppel, Moshe and Jonathan Schler (2004). "Authorship Verification as a One-class Classification Problem." In: *Proceedings of the Twenty-first International Conference on Machine Learning*, p. 62. DOI: 10.1145/1015330.1015448.

Moisl, Hermann (2011). "Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora." In: *Journal of Quantitative Linguistics* 18.1, pp. 23–52. DOI: 10.1080/09296174.2011.533588.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2008). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." In: *Political Analysis* 16.4, pp. 372–403. DOI: 10.1093/pan/mpn018.

Moretti, Franco (2013). *Distant Reading*. 1st edition. Verso.

Plecháč, Petr and Artjoms Šeļa (2021). "Applications." In: *Versification and Authorship Attribution*. Karolinum Press, pp. 69–91. DOI: 10.14712/9788024648903.5.

Reeve, Jonathan (2015). *Imperial Voices: Gender and Social Class among Shakespeare's Characters, a Stylometric Approach*. URL: https://jonreeve.com/2015/03/imperial-voices/.

Stammbach, Dominik, Maria Antoniak, and Elliott Ash (2022). "Heroes, Villains, and Victims, and GPT-3." In: *WNU 2022*, p. 47.

Sternberg, Meir (1982). "Proteus in Quotation-Land: Mimesis and the Forms of Reported Discourse." In: *Poetics Today* 3.2, pp. 107–156. DOI: 10.2307/1772069.

Székely, Gábor J. and Maria L. Rizzo (2013). "Energy Statistics: a Class of Statistics Based on Distances." In: *Journal of Statistical Planning and Inference* 143.8, pp. 1249–1272. DOI: 10.1016/j.jspi.2013.03.018.

Underwood, Ted, David Bamman, and Sabrina Lee (2018). "The Transformation of Gender in English-Language Fiction." In: *Journal of Cultural Analytics* 3.2, p. 11035. DOI: 10.22148/16.019.

Vishnubhotla, Krishnapriya, Adam Hammond, and Graeme Hirst (2019). "Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama." In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 29–34. DOI: 10.18653/v1/W19-2504.

Appendix

A Relatively-More-Frequent Words

French		German		Shakespeare	
Female	**Male**	**Female**	**Male**	**Female**	**Male**
vous	diable	ach	der	husband	the
époux	la	o	die	you	of
mère	ami	du	teufel	alas	this
amant	les	vater	und	love	sir
mari	parbleu	mutter	ein	husbands	and
tante	maître	er	des	me	we
hélas	morbleu	mich	in	romeo	king
coeur	des	liebe	den	lysander	our
rivale	amis	mama	kerl	willow	their
ne	morgué	papa	kaiser	pisanio	duke
malheureuse	serviteur	nein	ihr	sister	three
quil	belle	dat	euch	nerissa	her
mon	vin	mein	auf	yours	to
me	un	herz	dem	o	whom
maman	heureux	gemahl	wir	pray	lordship
frère	leur	gott	könig	mother	in
fâchée	rome	geliebter	sache	nurse	stand
père	peuple	kind	also	i	noble
obligée	boire	ihn	hm	woman	ha
sûre	soldats	lieber	majestät	malvolio	dog
dorante	peste	nicht	oder	prithee	certain
il	prêt	nich	volk	my	kate
soeur	rival	sie	euer	orlando	master
amour	dé	mann	das	boyet	sword
lui	sénat	weh	unter	do	follow
heureuse	ça	dir	im	false	soldiers
que	messieurs	dich	zum	ring	his
pleurs	coquin	mellefont	freund	emilia	caesar
cruel	gens	ja	krieg	refuse	us
lamour	du	ihm	durch	troilus	law
chevalier	allons	angst	hölle	pilgrim	friends
seule	beauté	freundin	zu	windsor	york
aimée	au	wat	gnaden	would	money
lingrat	lhomme	doch	wein	rosalind	pompey
valère	par	mamachen	heer	such	england
aime	obligé	so	mit	weep	present
aimer	lé	mir	bürger	faith	warwick
maime	bon	fritz	jeder	suit	great
hans	cents	arme	herren	am	heads
chère	bâton	gurli	rom	diana	ready
ingrat	quatre	lieb	land	never	business

B Bayesian Regression Models: Effect of Gender on Distinctiveness

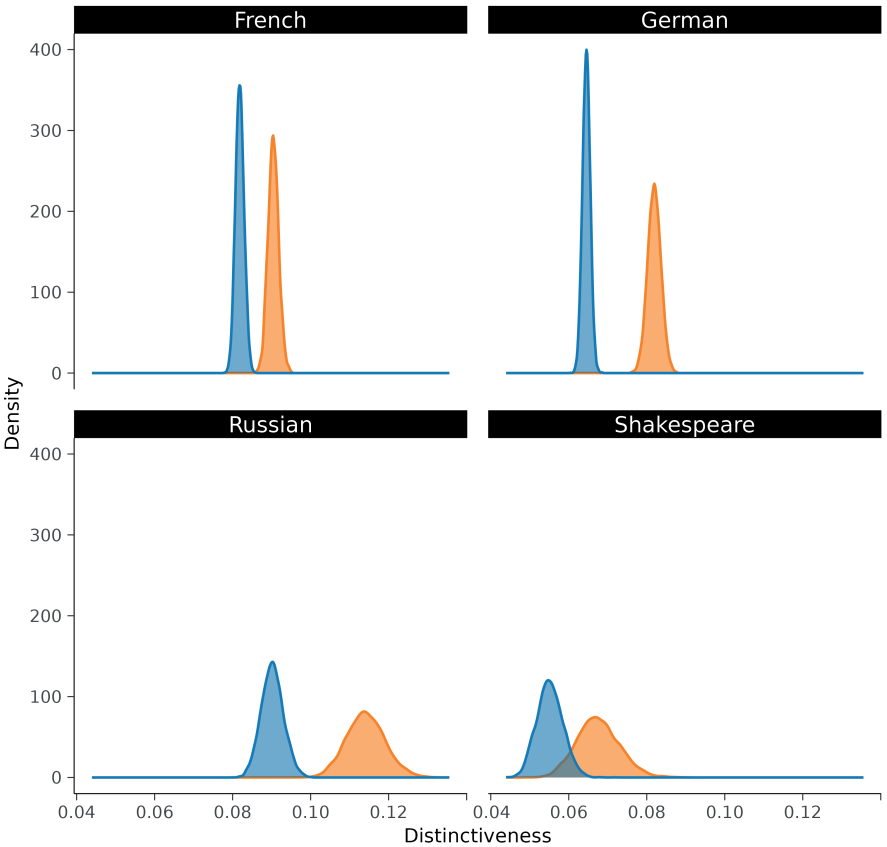


Fig. 4: Character distinctiveness, predicted from posterior, estimate of grand mean (no group-level effects), 6000 draws. Predictions are made for a counterfactual "median" character role, who has 20.9 % of dialogue share. Predictions are presented at natural scale.

Is the perceived gender effect 'real'? In technical terms, what is the direct influence of character gender (G) on distinctiveness scores (D) across traditions (T), conditioned on the share of dialogue they have (S)? To answer this, we fit a Bayesian multilevel multiple regression with group-level estimates for individual plays (P). We chose to model at the level of plays both because our D statistic is tied to the context of a single play, and because character features coming from the same

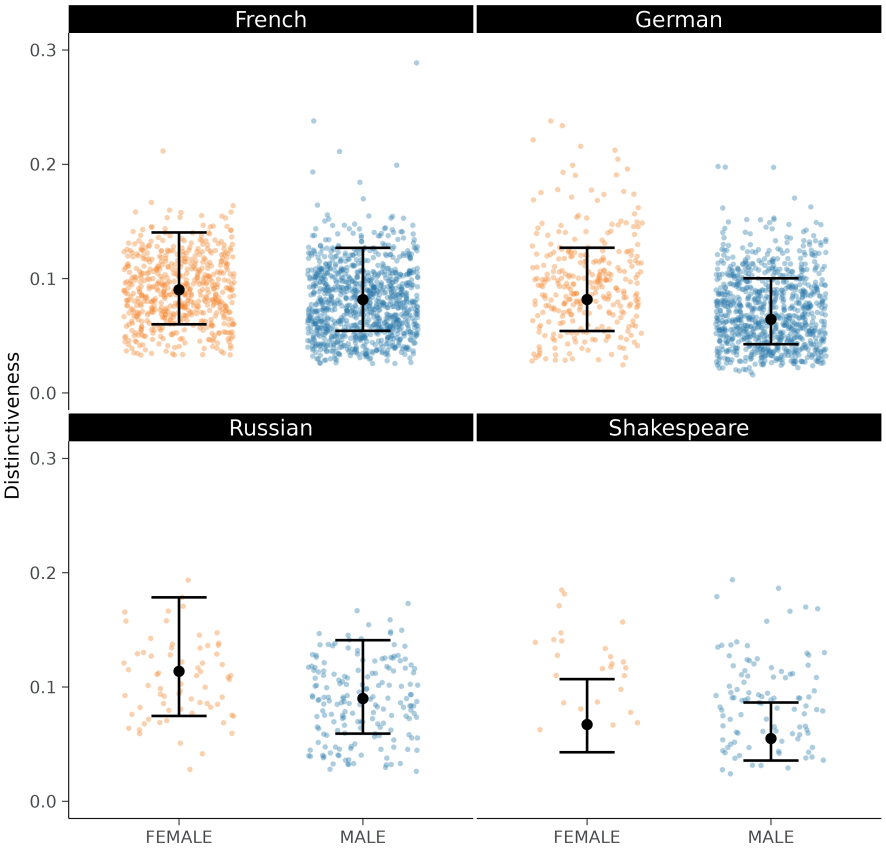


Fig. 5: Posterior predictions for gender, marginal of individual plays. Errorbars show .95 CI. Empirical data is plotted in colour, 5 extreme cases (>0.3) are filtered out. Predictions are presented at natural scale.

play are not independent (e.g. there cannot be two characters with 60 % of the dialogue). Modelling this way also significantly improved predictions. Gender is allowed to interact by corpus, yielding a single, cross-linguistic model that makes compatible predictions for different traditions. In brms formula syntax:

$$\log(D) \sim G * T + T * (S + I(S\hat{}2)) + (1|P) \tag{1}$$

Based on sample observation, we used a Gaussian prior for log-transformed D scores. We could have also fitted the original values, but D scores have extreme outliers that extend the tail: the model has much easier time with sampling and chain convergence on a log-transformed domain. We chose a quadratic term for S,

because the relationship between D and S is U-shaped. Importantly, 'unknown' gender entities are filtered, because often (but not always) this is not data that is missing, but entries that are incompatible with a binary classification:[1] primarily collective or compound entities (people, choirs, soldiers). It would have been possible to use standard strategies, like imputation, to 'repair' the data, but that approach would be incorrect.

Posterior estimates for distinctiveness by gender are shown in Figure 4. Based on the figure, we can be most confident about the difference in German and least confident in Shakespeare (few characters and, specifically, few women with large dialogue shares). The differences in means, however, appear consistent. As calculated from the posterior: in French, female characters are more distinctive by only .009 ($\pm$ .003); in German, by .017 ($\pm$ .003); in Russian by .023 ($\pm$ .009, the widest CI); and in Shakespeare by .012 ($\pm$ .008).

To understand the full extent of variation across different plays, it is useful to look at the marginal posterior means of the plays (Fig. 5). Here, the difference in distinctiveness between genders remains visible, but there is a better estimation of the global uncertainty and variation across different texts. Note that the confidence intervals in Fig. 5 are asymmetric (wider on the upper arm), having been transformed from symmetric intervals on a log domain.

1 In modern terms, it is vexing to be forced to reduce characters to a gender binary, but since gender non-conforming characters are virtually unrepresented in this predominantly historical corpus, the point is moot.