Gender Profiling for Slovene Twitter Communication: The Influence of Gender Marking, Content and Style

Ben Verhoeven
CLiPS Research Center
University of Antwerp
Prinsstraat 13, Antwerp, Belgium
ben.verhoeven@uantwerpen.be

Iza Škrjanec
Jožef Stefan International Postgraduate School
Jamova cesta 39, Ljubljana, Slovenia
skrjanec.iza@gmail.com

Senja Pollak
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
senja.pollak@ijs.si

Abstract

We present results of the first gender classification experiments on Slovene text to our knowledge. Inspired by the TwiSty corpus and experiments (Verhoeven et al., 2016), we employed the Janes corpus (Erjavec et al., 2016) and its gender annotations to perform gender classification experiments on Twitter text comparing a token-based and a lemma-based approach. We find that the token-based approach (92.6% accuracy), containing gender markings related to the author, outperforms the lemma-based approach by about 5%. Especially in the lemmatized version, we also observe stylistic and content-based differences in writing between men (e.g., more profane language, numerals and beer mentions) and women (e.g., more pronouns, emoticons and character flooding). Many of our findings corroborate previous research on other languages.

1 Introduction

Various computational linguistic and text mining tasks have so far been investigated for Slovene. Standard natural language processing (NLP) tools have been developed, such as preprocessing tools for lemmatization (Juršič et al., 2010), tagging (Grčar and Krek, 2012; Ljubešić and Erjavec, 2016) and parsing (Dobrovoljc et al., 2012), more recently adapted also for preprocessing non-standard texts, such as historical or computer-mediated Slovene (Ljubešić et al., 2016). However, not much attention has been paid to computational stylometry. While Zwitter Vitez (2013) applied authorship attribution, author profiling received nearly no attention. Recently Ljubešić and Fišer (2016) have addressed the classification of private and corporate Twitter accounts, while – to the best of our knowledge – we are the first to address gender profiling.

Author profiling is a well-established subfield of NLP with a thriving community gathering data, organizing shared tasks and publishing about this topic. Author profiling entails the prediction of an author profile – i.e., sociological and/or psychological characteristics of the author – based on the text that they have written. The most prominent author profiling task is gender classification, other tasks include the prediction of age, personality, region of origin, and mental health of an author.

Gender prediction became a mainstream research topic with the influential work by Koppel et al. (2002). Based on experiments on a subset of the British National Corpus, they found that women have a more relational writing style (e.g., using more pronouns) and men have a more informational writing style (e.g., using more determiners). Later gender prediction research remained focused on English, yet the attention quickly shifted to social media applications (Schler et al., 2006; Burger et al., 2011; Schwartz et al., 2013; Plank and Hovy, 2015). In the last few years, more languages have received attention in the context of author profiling (Peersman et al., 2011; Nguyen et al., 2013; Rangel et al., 2015; Rangel et al., 2016), with the publication of the TwiSty corpus containing gender information on Twitter authors for six languages (Verhoeven et al., 2016) as a highlight so far. We aim to contribute to the language diversity of this research line by looking at Slovene.

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 119–125,
Valencia, Spain, 4 April 2017. ©2017 Association for Computational Linguistics

Slovene belongs to languages with a pronounced morphology for gender. Nouns (and personal pronouns) have a defined grammatical gender (feminine, masculine, and neuter) in agreement with which other parts of speech can be inflected. Some of those structures allow for the identification of the author's gender in self-referring context. For example, the author's gender can be reflected in corresponding self-describing noun forms, e.g., *učitelj/učiteljica* (teacher$_{\text{male/fem}}$), and even more frequently in agreements of adjectives, e.g., *lep/lepa* (beautiful$_{\text{male/fem}}$), and non-finite verb forms, such as l-participles,[1] e.g., *sem delal/delala* (I worked$_{\text{male/fem}}$), which makes these markings a potentially useful feature for gender identification. As the inflected gender features might overshadow other relevant features, such as content and style, we investigate not only a token-based, but also a lemma-based approach. Disregarding easily manipulatable gender features (e.g., grammatical gender markings) can be seen as a first step towards an adversarial stylometry system, where we assume that the writer might not be who they claim to be. A second step would be to disregard content features, which can be easily manipulated as well. The lemma-based approach also allows for meaningful results to contribute to the field of sociolinguistics.

For our research in Slovene, findings in author profiling for related languages are of interest, especially with regard to feature construction due to morphological richness. Kapočiūtė-Dzikienė et al. (2015) predicted age and gender for Lithuanian literary texts. Lithuanian parliamentary texts were used to identify the speaker's age, gender and political view in Kapočiūtė-Dzikienė et al. (2014). A study of Russian showed there is a correlation between POS-bigrams and a person's gender and personality (Litvinova et al., 2015). Another relevant contribution to the field for Russian was the interdisciplinary approach to identifying the risk of self-destructive behavior (Litvinova and Litvinova, 2016). Experiments for gender identification for Russian show the advantages of grammatical features. Sboev et al. (2016) removed topical and genre cues from the corpus of picture descriptions and personal letters in Russian and ran tests for various features and machine learning algorithms to find the combination of grammatical information (POS-tags, noun case, verb form, gender, and number) and neural networks performed best. As far as we know, no gender classification of tweets in these languages has been presented.

The present paper is structured as follows: in Section 2, we describe the Janes Tweet corpus and its modification for the experiments, which are presented in Section 3. In Section 4, we discuss the results in terms of performance and feature interpretation, while in Section 5 we conclude our study and propose further work.

2 Corpus Description

For our experiments, the Janes corpus (Erjavec et al., 2016; Fišer et al., 2016) of user-generated Slovene was adapted to match the TwiSty corpus setting (Verhoeven et al., 2016). We will first introduce the Slovene source corpus and then describe our reformatting of it for the current research.

The Janes corpus was collected within the Janes national research project[2] and consists of documents in five genres: tweets, forum posts, news comments, blog entries, and Wikipedia user and talk pages. The Twitter subcorpus is the largest Janes subcorpus. The tweets were collected using the TweetCat tool (Ljubešić et al., 2014), which was designed for building Twitter corpora of smaller languages. Employing the Twitter Search API and a set of seed terms, the tool identifies users writing in the chosen language together with their friends and followers. The tool outputs tweets together with their metadata (tweet ID, time of creation and retrieval, favorite count, retweet count, and handle). In total, the corpus includes tweets by 8,749 authors with an average of 850 tweets per author.

The authors were manually annotated for their gender (female, male and unknown) and account type (private and corporate). Personal accounts are considered as private account types, while companies and institutions count as corporate ones. The gender tag was ascribed based on the screen name, profile picture, self-description ('bio') and – in the few cases that this was not sufficient – the use of gender markings when referring to themselves. The account type was annotated given the user name, self-description and (typically impersonal) content of tweets. Since the focus of our study

[1] Verb l-participles is the name for the Slovene participles that end in letter 'l' in the masculine form and can be used for past, future and conditional constructions.

[2] http://nl.ijs.si/janes/

	WRB	MAJ	Accuracy	Precision	Recall	F1-score
Token	56.9	68.5	92.6	92.7	92.6	92.6
Lemma	56.9	68.5	87.9	87.9	87.9	87.9

Table 1: Results of gender prediction experiments based on tokenized text and on lemmas. Abbreviations: WRB = Weighted Random Baseline, MAJ = Majority Baseline. Precision, Recall and F1-score are averaged over both classes (since both classes matter).

was the binary prediction of female or male gender, only private male and female accounts were considered in the experiments.

Given the multilingual context of user-generated content, each tweet had to undergo language identification. For this the `langid.py` program (Lui and Baldwin, 2012) was used. The identified language tags were additionally corrected with heuristics resulting in four possible tags for the entire corpus: Slovene, English, Serbian/Croatian/Bosnian, and undefined (Fišer et al., 2016).

This subcorpus of Janes was reformatted to resemble the TwiSty corpus in order to address the same task of author profiling. There are however a few differences that we should mention for completeness. The Janes corpus does not have the personality type information available for the users and the language identification was performed in a different way.

3 Experiments

The experimental setup of this research is largely based on the TwiSty experiments (Verhoeven et al., 2016). We will briefly describe this approach and explain our additions.

First of all, to ensure comparability of instances, we construct one instance per author by concatenating 200 language-confirmed tweets. Authors with less than 200 tweets are discarded. All user mentions, hashtags and URLs were anonymized by replacing them with a placeholder token to abstract over different instances to a more general pattern of their use. The final dataset contains 3,490 instances with more men (68.5%) than women (31.5%), see Table 2.

The gender prediction task is set up as a two-class classification problem with classes *male* and *female* in a standard tenfold cross-validation experiment using the LinearSVC algorithm in `scikit-learn` (Pedregosa et al., 2011). We used n-gram features on both word ($n = [1, 2]$) and character ($n = [3, 4]$) level. We did not perform any feature selection, feature weighting or parameter optimization.

	Count	Percentage
Male	2,391	68.5
Female	1,099	31.5
Total	3,490	100

Table 2: Corpus statistics: male and female private Twitter users represented by 200 tweets per author.

The experiment was performed in two different settings: on tokenized text,[3] and on lemmatized text. The lemmatized text is available in the Janes corpus (for lemmatization process see Ljubešić and Erjavec (2016)). The results of these experiments can be found in Table 1 and will be discussed in Section 4.

We also performed the experiment on a normalized version of the text that was available in the Janes corpus. This means that substandard spellings were corrected to the standard form, especially including the restoration of diacritics. Our expectation was that standardizing the text would allow for 1) certain features to cluster together and get stronger and thus more generalizable; and 2) disambiguation of certain words due to diacritics restoration. However, the results of this experiment were near-identical to the experiment on tokenized text, so we will not further discuss this here.

4 Discussion

Our experiments show a very high and interpretable result. Using tokenized text clearly outperforms the use of lemmas by around 5%, but both systems appear to work really well, significantly outperforming both the weighted random baseline (WRB) and majority baseline (MAJ).

Interestingly, our results are higher than the state-of-the-art results for the different languages

[3] Using the `happierfuntokenizing` script by Christopher Potts (`http://wwwbp.org`), as also used by Verhoeven et al. (2016).

in TwiSty. The most comparable language in data size would be Portuguese, which achieves 87.6%, while we achieve 92.6% for Slovene. As our feature analysis below will show, the difference lies in the gender markings.

Slovene encodes gender more extensively than Romance languages do. Especially the frequently used verb l-participles are important features for gender profiling, because a gender marking for the author is present every time the author is the subject of the past and future tense and conditional verb mood that are expressed by the auxiliary and the participle. Although agreement is partly informative also in other Romance languages, i.e., through participle agreement in French, e.g., *je suis allé/allée* (I went$_{\text{male/fem}}$), Italian, e.g., *io sono andato/andata* (I went$_{\text{male/fem}}$), Spanish, e.g., *yo fui invitado/invitada* (I was invited$_{\text{male/fem}}$), or adjectival agreement in French, e.g., *je suis heureux/heureuse* (I am happy$_{\text{male/fem}}$) or Spanish, e.g., *yo soy viejo/vieja* (I am old$_{\text{male/fem}}$), the gender markings are much less frequent than in Slavic languages, such as Slovene.

By lemmatizing the text, we remove this effect and we observe the performance of the system to lower to 87.9% which is very comparable to that of Portuguese and Spanish in the TwiSty paper (Verhoeven et al., 2016).

We also investigated the most informative features that `scikit-learn` outputs when retraining the model on the entire dataset (i.e., no tenfold). We extracted a ranked list of the 1,000 most informative features per class[4] and were able to make a comparison between the genders and between the token- and lemma-based approaches.

The most informative features of the token-based approach confirm very clearly our explanation of the higher performance of this approach compared with the lemma-based approach. The bulk of the most informative features can be related to gender markings on verb l-participles (e.g., MALE: *mislil* (thought), *bil* (been), *vedel* (known), *gledal* (watched); FEMALE: *mislila* (thought), *dobila* (gotten), *rekla* (said), *videla* (seen)), as well as feminine adjective forms (e.g., *ponosna* (proud), *vesela* (happy)).

The informative features for the lemma-based approach contain almost no gender markings. However, many interesting stylistic and content-based features become apparent, some of them also occurring lower on the ranking with the token-based approach.

We found several word and character features associated with the use of profane language that are strongly linked to the male category, e.g., *jebati* and *fukati* (to fuck), *pizda* and *pička* (cunt), *rit* (ass), *srati* (to shit), *kurec* (dick), *joške* (boobs). Another characteristic distinctive of the male class is non-alphabetical symbols including symbols for euro (€) and percent (%), and numerals (as digits) – the latter were also found to be more indicative of male authors and speakers in an English corpus of various genres (Newman et al., 2008) and the spoken part of BNC (Baker, 2014). Interestingly, vulgar expressions do not occur among the most informative features of the female category, while a small number of numerals can be found. The female category is distinguished by the use of emoticons (*;3*, *:**, *:)*, ♥), however the emoticon with tongue (*:P*) is related to the male category. Among the most informative features on both lemma- and token-level various interjections often combined with character flooding occur in the female category: *(o)joj* (oh), *oh* (oh), *ah* (oh), *ha* (ha), *bravo*, *omg*, *jaaa* (yaaas), *aaa* (argh), *ooo* (oooh), *iii* (aaaw). The female category further displays linguistic expressiveness in intensifiers (*ful* (very), *čist* (totally)) and adjectives and adverbs denoting attitude (*grozen* (horrible), *lušten* (cute), *gnil* (rotten), *čuden* (weird)), but these require further support in analysis.

A strong stylistic feature of the female category is referring to self with personal and possessive pronouns in first person: *jaz* (me), *zame* (for me), *moj* (my/mine) on the lemma-level, and *meni* (to me), *moje* (my/mine), *mene* (me$_{\text{accusative}}$) on the token-level with some of these features on both levels occurring within word bigrams (*biti_moj* for be_mine). Referring to others is also more present in the female category, namely with possessive pronouns for third person singular (*njen* (her/hers), *njegov* (his)) and first person plural (*naš* (our/ours)). This corroborates prior findings for English where women also use more pronouns than men (Schler et al., 2006).

A minor feature that requires further analysis is the use of diminutive endings in the female category (*-ček* and *-kica*).

The lemma-based approach provides insight into interesting tendencies regarding the content.

[4] These lists are available online at: `https://github.com/verhoevenben/slovene-twisty`.

The topics in the male category are associated with drinking (*pivo*/*pir* (beer), *bar*; *piti* (to drink) in the token-based list), sports (*tekma* (game), *šport* (sports), *fuzbal* (football), *zmaga* (win)) and motoring (*guma* (tire), *avto* (car), *voziti* (to drive/ride)). In the female category, a topic on food and beverages is also present, but with a different focus (*hrana* (food), *čaj* (tea), *čokolada* (chocolate), *sladoled* (ice cream)). Both female and male authors refer to other people, but they focus on different agents. Referring to women (*ženska*), men (*moški*), kinship (*starš* (parent), *mami* (mom), *otrok* (child), *babica* (grandma), *teta* (aunt)), female friends (*prijateljica*) and female colleagues (*kolegica*) relates more with the female category, while we can find references to wives (*žena*), male colleagues (*kolega*) and male friends (*prijatelj*) in the male category.

The token- and lemma-based levels of both categories display various modality markers: *marati* (to like), *ne_moči* (not_able), *zagotovo* (definitely), *želim* (I wish) for the female category, and *rad* (like/want$_{\text{male}}$), *verjetno* (probably), *hotel* (wanted$_{\text{male}}$), *želel* (wished$_{\text{male}}$), *potrebno* (necessary) for the male category.

It is interesting to note that these stereotype-confirming gendered features strongly resemble earlier results on social media data for English. In their research on Facebook text, Schwartz et al. (2013) also found men to use more swear words and women to use more emoticons. Similarly, according to a study by Bamman et al. (2014) on English tweets, emoticons and character flooding are associated with female authors, while swear words mark tweets by male authors. Again, both groups use kinship terms, but with a divergence similar to our finding.

5 Conclusions and Further Work

We conclude that the classification of Twitter text by gender works very well for Slovene, especially when the system can use the gender inflection on the verb l-participles, but also in a lemmatized form where the system can use stylistic and content features.

Should one wish to use gender classification in an adversarial setting – i.e., when you take into account people trying to actively mislead a reader by posing as a different person or gender – the content features should also be removed from the experiment as they too can be easily manipulated. Function words and POS-tags are the best features in this setting, as they are not under conscious control (Pennebaker, 2011). Slovene would be an interesting language to research this for, as pronouns – which are considered to be very salient author profiling features – are often not explicit.

Acknowledgements

The work described in this paper was partially funded by the Slovenian Research Agency within the national basic research project Resources, Tools and Methods for the Research of Nonstandard Internet Slovene (J6-6842, 2014-2017). The first author is supported by a PhD scholarship from the FWO Research Foundation – Flanders.

References

Paul Baker. 2014. *Using Corpora to Analyze Gender*. Bloomsbury, London.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kaja Dobrovoljc, Simon Krek, and Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. In Tomaž Erjavec and Jerneja Ž. Gros, editors, *Zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C*, pages 42–47. Institut Jožef Stefan, October.

Tomaž Erjavec, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Darja Fišer. 2016. Gold-standard datasets for annotation of Slovene computer-mediated communication. In *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2016)*. Brno, Češka.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2016. Janes v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2):67–99.

Miha Grčar and Simon Krek. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec and J. Žganec Gros, editors, *Proceedings of the 8th Language Technologies Conference*, volume C, pages 89–94, Ljubljana, Slovenia, October. IJS.

Matjaz Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. LemmaGen: Multilingual lemmatisation with induced ripple-down rules. *J. UCS*, 16(9):1190–1214.

Jurgita Kapočiūtė-Dzikienė, Ligita Šarkutė, and Andrius Utka. 2014. Automatic author profiling of Lithuanian parliamentary speeches: Exploring the influence of features and dataset sizes. In *Human Language Technologies The Baltic Perspective, Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania.

Jurgita Kapočiūtė-Dzikienė, Andrius Utka, and Ligita Šarkutė. 2015. Authorship attribution and author profiling of Lithuanian literary texts. In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*. Hissar, Bulgaria.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Tatiana Litvinova and Olga Litvinova. 2016. Authorship profiling in Russian-language texts. In *Proceedings of the 13th International Conference on Statistical Analysis of Textual Data (JADT)*. Nice, France.

Tatiana Litvinova, Pavel Seredin, and Olga Litvinova. 2015. Using part-of-speech sequences frequencies in a text to predict author personality: a corpus study. *Indian Journal of Science and Technology*, 8(9):93–97.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Nikola Ljubešić and Darja Fišer. 2016. Private or corporate? Predicting user types on Twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 38–46.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2014. Tweetcat: A tool for building Twitter corpora of smaller languages. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. ELRA, Reykjavik, Iceland.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Slovene data : historical texts vs. user-generated content. In Heike Zinsmeister Stefanie Dipper, Friedrich NeuBarth, editor, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pages 146–155.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*. Jeju, Korea. ACL.

Matthew Newman, Carla Groom, Lori Handelman, and James Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211236.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, Theo Meder, and C-M Au Yeung. 2013. TweetGenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter*, 4(4).

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, M. Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

James W. Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter -or- how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. Lisbon, Portugal.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Working Notes*. CEUR.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *CLEF 2016 Working Notes*. CEUR-WS.org.

Aleksandr Sboev, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, and Ivan Moloshnikov. 2016. Machine learning models of text categorization by author gender using topic-independent features. In *Proceedings of the 5th International Young Scientist Conference on Computational Science*. Procedia Computer Science, Krakow, Poland.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9).

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: a multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. ELRA, Portorož, Slovenia.

Ana Zwitter Vitez. 2013. Le décryptage de l'auteur anonyme : l'affaire des électeurs en survêtements. *Linguistica*, 53(1):91–101.