Explanation in Computational Stylometry

Walter Daelemans

CLiPS, University of Antwerp, Belgium,
walter.daelemans@ua.ac.be

Abstract. Computational stylometry, as in authorship attribution or profiling, has a large potential for applications in diverse areas: literary science, forensics, language psychology, sociolinguistics, even medical diagnosis. Yet, many of the basic research questions of this field are not studied systematically or even at all. In this paper we will go into these problems, and suggest that a reinterpretation of current and historical methods in the framework and methodology of machine learning of natural language processing would be helpful. We also argue for more attention in research for explanation in computational stylometry as opposed to purely quantitative evaluation measures and propose a strategy for data collection and analysis for achieving progress in computational stylometry. We also introduce a fairly new application of computational stylometry in internet security.

1 Meta-knowledge Extraction from Text

The form of a text is determined by many factors. Content plays a role (the topic of a text determines in part its vocabulary), text type (genre, register) is important and will determine part of the writing style, but also psychological and sociological aspects of the author of the text will be sources of stylistic language variation. These psychological factors include personality, mental health, and being a native speaker or not; sociological factors include age, gender, education level, and region of language acquisition.

Writing style is a combination of consistent decisions in language production at different linguistic levels (lexical choice, syntactic structures, discourse coherence, ...) that is linked to specific authors or author groups such as male authors or teenage authors. It remains to be seen whether this link is consistent over time and whether there are style features that are unconscious and cannot be controlled, as some researchers have argued. The basic research question for computational stylometry seems then to describe and *explain* the causal relations between psychological and sociological properties of authors on the one hand, and their writing style on the other. These theories can be used to develop systems that generate text in a particular style, or perhaps more usefully, systems that detect the identity of authors (authorship attribution and verification) or some of their psychological or sociological properties (profiling) from text.

A limit hypothesis arising from this definition is that style is unique for an individual, like her fingerprint, earprint or genome. This has been called the *human stylome hypothesis*:

> '(...) authors can be distinguished by measuring specific properties of their writings, their stylome as it were.' [1]

Reliable authorship attribution and profiling is potentially useful in many areas: literary science, sociolinguistic research, language psychology, social psychology, forensics, medical diagnosis (detecting schizophrenia and Alzheimer's), and many others. In Sect. 3 we describe results in the context of an internet security case study as an example of useful computational stylometry. However, the current state of the art in computational stylometry seems not advanced enough to always guarantee the levels of reliability expected.

There are many excellent introductions to modern computational methods in stylometry [2–5] describing the methods and feature types used. Feature types include simple character n-grams, punctuation, token n-grams, semantic and syntactic class distributions and patterns, parse trees, complexity and vocabulary richness measures, and even discourse features.

Computational stylometry should be investigated in a Natural Language Processing (NLP) framework, more specifically as one of three levels of text understanding. The goal of text understanding is to extract knowledge from text and present it in a reusable format. NLP has seen significant progress in the last decade thanks to a switch to statistical and machine learning based methods in research and increased interest because of commercial applicability (Apple's SIRI and Google translate are only two examples of recent high impact commercial applications of NLP). The three types of knowledge we distinguish that can be extracted from text are: (i) objective knowledge (answering the who, what, where, when, ... questions), (ii) subjective knowledge (who has which opinion about what?), and (iii) metaknowledge (what can we extract about the text apart from its contents, mainly about its author?). Computational stylometry belongs in the latter category.

Core research in NLP addresses the extraction of objective knowledge from text: which concepts, attributes, and relations between concepts can be extracted from text, including specific relations such as causal, spatial and temporal ones. Research is starting also on the Machine Reading loop (how to use background knowledge in text analysis and conversely how to build up background knowledge from text). See work on Watson for state of the art research at this first level [6]. In addition to the extraction of objective knowledge, the large amount of text produced in social networks has motivated research to focus also on the extraction of subjective knowledge (sentiment and opinion). Never before have so many non-professional writers produced so much text, most of it subjective and opinionated (reviews, blogs, e-mail, chat, ...) [7]. Extraction of *metaknowledge* is conceptually a different type of knowledge extraction from text than the other two types. Where objective and subjective knowledge extraction try to make explicit and structure knowledge that is present in unstructured textual information, metaknowledge concerns knowledge about the author of the text (psychological and sociological properties, and ultimately identity), so outside the text. Recent advances in knowledge extraction from text at all these three levels have been made possible thanks to the development of robust and fairly

accurate text analysis pipelines for at least some languages. These pipelines make possible the three types of knowledge extraction described earlier thanks to morphological analyzers, syntactic parsers, sentence semantics (including semantic roles and the analysis of negation and modality), and discourse processing (e.g. coreference resolution). Of course, the point is that by integrating in this process also analyses from objective and subjective knoweldge extraction, more interesting theories about the extraction of metaknowledge become possible in principle.

For all types of knowledge extraction, supervised machine learning methods have been a powerful solution. Based on annotated corpora, various properties of text are encoded in feature vectors, associated with output classes, and machine learning methods are used to learn models that generalize to new data. It is surprising that much computational stylometry research is still explicitly linked to the idea of automatic text categorization [8] (as used in document filtering and routing applications) rather than to supervised machine learning of language in general (unsupervised and semi-supervised learning methods will not be discussed here). It makes sense to treat computational stylometry within the same methodological paradigm as other knowledge extraction from text tasks. For example, making a distinctions between similarity-based methods and machine learning methods as in [9] is unproductive as the former is a type of machine learning method as well (lazy learning as opposed to eager learning) [10]. All techniques proposed before in the long history of stylometry can be reinterpreted as machine learning methods to our advantage. A good example of this is Burrow's delta which through its reinterpretation as memory-based learning [11] leads to increased understanding of the method and to new useful variations. It would be equally productive if new methods like *unmasking* [12] and variants would be framed as instances of stacked classifiers and ensemble learning, which they are, thereby providing more clarity.

In a supervised machine learning approach to computational stylometry we have to consider the features to be used to describe our objects of interest (complete texts or text fragments), feature selection, weighting and construction methods, machine learning algorithm optimization, and the usefulness of techniques like ensemble methods, active learning, joint learning, structured learning, one-class learning etc. We can also rely on proven evaluation methods and methodological principles for comparing features and methods. Systematic studies in such a framework will go a long way in coming up to Rudman's [14, 13] criticism that after more than 40 years of research and almost a thousand papers (many more counting conference contributions), modern authorship studies "have not yet passed a 'shake-down' phase and entered one marked by solid, scientific, and steadily progressing studies."

2 Problems in Computational Stylometry

Computational stylometry is an exciting field with a promise of many useful applications, but initial successes have underplayed the importance of many

remaining problems. So far, we already have encountered a number of unsolved basic research questions we will not go into in this paper, but that deserve more systematic study.

- Is style invariant or does it change with age and language experience? There is some work in this area (see [15] for an overview), but no large-scale systematic studies. If individual style changes over time, which seems to be the case, this is a confounding factor for attribution.
- Is style largely unconscious or can it be imitated? Again, there is some work on *adversarial stylometry*, but not enough for clear conclusions. Initial work [16] is not optimistic and shows that obfuscation reduces authorship identification methods to random behaviour.

Unless style markers can be found that are robust to aging and conscious manipulation, the human stylome hypothesis should be discarded. But there are other problems that need urgent attention as well.

2.1 Scalability and Character n-grams

Another problem that has only relatively recently received attention is the issue of *scalability*. Authorship attribution and profiling work reasonably well when large amounts of text are available, and in the case of authorship attribution, few candidate authors for an unattributed text are present, one of which is the author (the closed case). This model fits literary disputed authorship cases with a small set of candidate authors, for example. In more realistic situations, we have short texts (for example letters or e-mails), and many potential authors. In [17], we showed, using a corpus of same-topic essays by 145 different authors, that with many potential authors or with short texts, attribution accuracy quickly decreases to levels that are still above baseline but nevertheless too low for practical applications. We also saw that simple character n-grams are more scalable than more complex (lexical and syntactic) feature sets. More work on scalability has been done (with better reported results) in [9]. The same scalability issues apply to profiling applications in computational stylometry as well.

The superiority of character n-grams is something which is often attested in stylometry: character n-grams often outperform more complex feature sets [18]. There is a good reason for this. They provide an excellent tradeoff between sparseness and information content. Because of their higher frequency compared to other feature types such as tokens, better probability estimates are possible for character n-grams, while at the same time they combine information about punctuation, morphology (character n-grams can represent morphemes as well as roots), lexicon (function words are often short), and even context (when extracting n-grams at sentence level rather than at token level). In addition they are tolerant to spelling variation and errors. On top of that, from a practical point of view, models based on character n-grams are very easy to construct and they are language-independent. There may also be a more negative explanation for their success in computational stylometry: it may be the case that the

language processing tools that have to provide the more sophisticated linguistic analysis are not accurate enough and generate too much noise in the document representations.

The supervised machine learning context also helps us in understanding that scalable authorship attribution should not be framed as a multi-class learning problem, but as a binary or even one-class learning problem[19]. The real problem in authorship attribution is not to decide who from a limited number of authors, for all of whom we have training material, has written a particular text (the closed case), but to decide whether the new text was written by a particular author (for whom we have training material), or not, a task known as *authorship verification* (the open case). Very recently, this was defined in [20] as the fundamental problem of authorship attribution:

> 'Given two (possibly short) documents, determine if they were written by a single author or not.'

We will return to their solution in Sect. 2.2.

Successes with the closed case have lead to overoptimistic ideas about the possibilities of computational stylometry because of overfitting. When learning a model to distinguish between two or a few authors, there is no guarantee that the predictive features selected by the model will generalize to distinguishing from additional authors. Compare it to a fruit classification application: color will be a great feature to distinguish between apple and banana, but as soon as lemon and pear are added to the task, the model breaks down.

The human stylome hypothesis is trivially correct: given an unlimited supply of text from each person speaking a language, some combination of features can probably be found that uniquely discriminates anyone from all others. But we expect a stylome of an author to consist of a limited combination of features that are frequent enough to be found in all text written by that author so that generalization is possible.

2.2 Cross-genre stylometry

One of the most basic problems to be solved for computational stylometry is finding out how style, content, and genre interact in the generation of style. A straightforward strategy for avoiding topic detection rather than style detection is to exclude content words as features. However, topic words can be predictive as well (e.g. consistent selection of one word from a set of synonyms by authors or groups of authors). Although there is some work in this area (see for example chapter 4 in [21] and references therein), more systematic research is needed.

An even less researched aspect of computational stylometry is the effect of genre on attribution. To which extent do stylistic properties of individual authors or groups of authors transfer from one genre to the other? Can we expect that a model trained on essays written by someone will be able to identify his suicide note or blackmail letter? Again this is a well known problem in machine learning for the case where training and test data come from different distributions (the

domain adaptation problem [22]). Domain adaptation problems exist both for genre and for topic (in the case where features based on content words are used).

In a recent study [23] we tackled both the problem of verification (rather than attribution, i.e. the open case) and the problem of cross-genre generalization. As machine learning method we tested the "unmasking" technique, recently proposed [12] and well-received [24]. Suppose we want to verify that a text X with unknown authorship was written by the author of a text A. We could split both texts in chunks, and train a classifier to distinguish between both. If the resulting classifier turns out to have low generalization accuracy, X and A were probably written by the same author, if it turns out to be easy to distinguish then not. The approach turned out not to work very well because a limited number of features can wrongfully maximize the differences in writing style between two texts written by the same author. As a solution, [12] proposed a stacked classifier approach, in which a new classifier is built on the basis of a previous classifier by removing those features that are most discriminative between the two texts. The degradation curves that can be attested by applying these subsequent classifiers to the task are indicative of whether the two texts were written by the same author. In the case of a few features being responsible for most differences (same author), the degradation curve would fall quickly. In the case of many features being responsible for the differences (different authors), the drop is less dramatic. It has been attested that the approach works well for longer texts and for related tasks such as intrinsic plagiarism detection, but not for shorter texts below 10,000 words in size [25]. We tested whether the approach works for the cross-genre authorship verification task in the expectation that the genre markers would be limited and superficial and would therefore be among the first to be discarded in the unmasking approach, leading to a clear degradation curve indicative of same authorship. We refer to the paper [23] for a detailed description of the operationalization of the unmasking approach to our cross-genre case. We applied the approach to theatre and prose texts of five authors. Whereas for the within-genre case the approach worked as expected, it didn't work very well for the cross-genre case. Although some of the most discriminative features discarded were indeed genre-related (names of principal characters, stage directions, colloquialisms, ...), the approach did not hold. Further research with optimization of the many parameters in the approach is still needed, but it seems clear that we will need new methods for coping with cross-genre cases.

In conclusion, we have argued that many of the basic problems in computational stylometry are not being investigated at all or not sufficiently systematically. Good features for authorship verification and profiling should be robust against genre variation, topic variation, individual style change over time, and conscious manipulation. Methods should also be scalable to short texts. Arguably, it is the feature selection (or feature construction) problem which is most important in this field rather than the choice of machine learning method, although the specific problem of authorship verification may call for ensemble methods such as unmasking. But overall, what is lacking is explanation.

2.3 Explanation

One aspect of current machine learning of NLP research that the field of computational stylometry should not adopt is its unidimensional focus on quantitative evaluation. The goal of research should be to increase understanding rather than maximizing performance (which is an engineering criterion). In profiling, the field started in an excellent way regarding explanation with the gender assignment studies of [26]. They provided a plausible explanation for their success in distinguishing male from female authors in written text by hypothesising that women use more relational language, and men more informative (descriptive) language. That men are prone to more descriptive language use is reflected in text by a more frequent use of nouns, determiners, prepositions etc. Figure 1 shows some similar frequent features (part of speech tags, Pennebaker LIWC classes, tokens) related to male and female language use in Dutch. A darker colour under male or female indicates more frequent use. The hypothesis "men use more descriptive language" then explains a number of (correlated) lower level text features, and provides *insight* into how male and female gender is realized into text.

Fig. 1. Frequent Feature Types correlated with gender in Dutch.

Unfortunately, examples like this are rare. More frequently, a study will provide some new best result on a benchmark dataset using some clever feature engineering or classifier optimization, without attempting to provide an explanation for the results in a broader framework. At best there is some superficial error analysis. The current focus on challenges (also called shared tasks) using hastily compiled low quality "benchmark datasets" is an important culprit for this. There is seldom time for intelligent reflection on the construction of the datasets and the interpretation of the results, and there are no prizes for explanation, only for achieving the highest accuracy.

It could be argued that what is especially needed for improving understanding and explanation is (for each language) a real reference corpus which is carefully balanced according to genre, topic, age, and gender (and if possible also other psychological and sociological properties of the authors). Only then can real progress be made in solving the fundamental problems of computational stylometry. If we take the human stylome seriously as a hypothesis, we should start doing stylome-wide association studies (in analogy to genome-wide association studies) associating linguistic properties with author traits, and inferring explanatory concepts from the bottom-up interpretation of correlated sets of features. As in genetic studies, population stratification (i.e. balanced corpora) is a necessary precondition in such studies.

3 Detecting Harmful Content in Social Media

In a recently started cooperative Flemish project AMiCA[1], our goal is to identify possibly threatening situations (especially for children and adolescents) in social networks sites (SNS) by means of text and image analysis. The three critical situations targeted are cyberbullying, sexually transgressive behavior, and depression and suicidal behavior. For text-based analysis we see these tasks partly as instances of computational stylometry. For example, for the detection of transgressive behavior by pedophiles[2] it is important not only to be able to detect the typical grooming stages in pedophile behavior, but also to be able to detect age and gender of the text in order to check the information provided in the SNS profiles. For detecting suicidal emotions and insults in cyberbullying, similar computational stylometry tasks can be defined. Some early results of our team can be found in [28, 27, 29].

For the detection of pedophiles in SNS we have available some data from the Belgian SNS *Netlog* in the form of interaction with associated profile information (age, gender, and location). The data is challenging because the utterances are short and written in chat language which has properties completely different from standard language. The properties of chat language are based on the fact that the interactions should be quick and informal (spoken language like). This leads to omission of letters and words, abbreviations, acronyms, non-verbal and

[1] http://www.amicaproject.be

[2] A preparatory PhD project, DAPHNE, about this was started before AMiCA. See http://www.clips.ua.ac.be/projects/daphne

suprasegmental mimicry (for example character flooding, concatenation and even merger of words, emoticons), and many other strange phenomena. Investigating this, we found interesting reflections in our data of claims about sociolinguistic language variation in spoken language. For example, Fig. 2 from a submitted paper shows how much different chat language is from standard language for different age groups, genders and regions in the chat data. It is clear from this data that non-standard language use in chat is especially a property of adolescents, and that in their twenties, chatters revert to more standard language. Also, some attested facts about sociolinguistic variation in the Flemish Dutch region can be clearly shown in this data: for example that men use more non-standard language than women, that Western Flanders uses more non-standard language than other regions and so on.

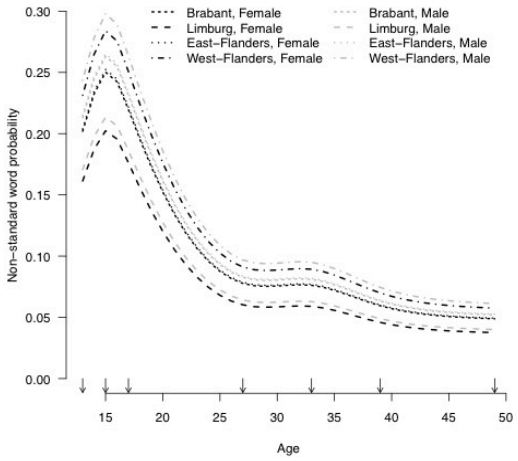

Fig. 2. Use of non-standard language in chat by Flemish sociological groups.

More important for our purposes is that this data can be used to train accurate classifiers for assigning age and gender. Our strategy is to develop two classifiers, one based on age and gender to check for mismatches between profiles provided and information extracted from the text of the interactions, and a second one to detect grooming behaviour, which can be detected to some extent by typical types of language use, for example directive language, and specific topics, for example 'coast is clear' checks.

In Fig. 3, the architecture we are working on is given.

By optimizing the classifiers for legally relevant age groups (minus 16 and plus 21 for example), very high f-scores (in the nineties) can be reached. Incidentally, this data is one example of a task where n-grams don't do very well.

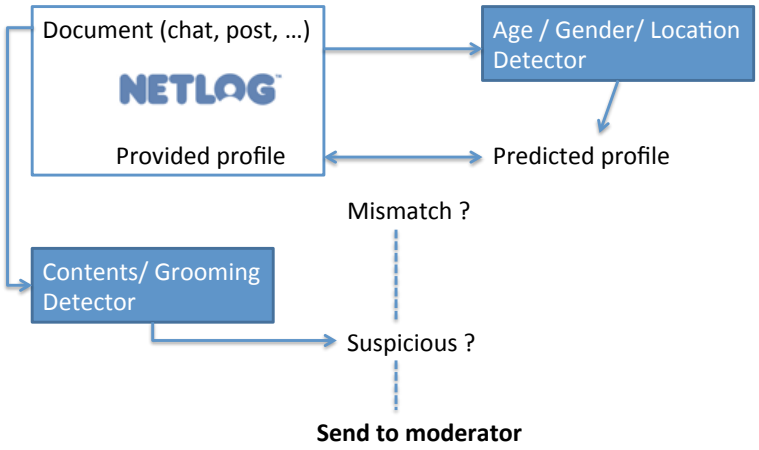

Fig. 3. Architecture for a pedophile detection system.

Unfortunately, because of the non-standard characteristics of the text, standard language text analysis tools cannot be used, so that we had to restrict ourselves to word tokens in these experiments. Current work on automatic normalization of chat language should make additional levels of analysis available soon.

Increasingly, the field has become interested in these more peripheral applications of computational stylometry. For example, in the context of CLEF, a shared task was organized in 2012 on pedophile detection[3]. There were many participating systems and some good very good detection results. However, the event illustrates many of the problems with data collection in shared tasks alluded to earlier. By collecting negative and positive data from different sources (the perverted justice website for the positive data and unrelated sources for the negative data), the task turns out to be artificially easy and generalization to other datasets very low. Also in this case, more work should be done on population stratification.

4 Conclusion

With our case study in guarding security of children and adolescents in SNS, we hope to have shown that computational stylometry has large application possibilities and is, thanks to advances in Natural Language Processing and Machine Learning, in a state where useful applications are already possible. But many fundamental problems of computational stylometry remain unsolved or even largely ignored. We are looking not just for a system that reaches a certain target accuracy in a task, but for explanations, and for systems that are scalable,

[3] http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html

and that generalize over different genres and topics in their author identification and profiling results. It seems clear that a systematic study of the components and concepts of style will only be possible by collecting a large balanced dataset for each language of a type that doesn't yet exist in current benchmark efforts.

Acknowledgements

I gratefully acknowledge the support of various research funds, most notably FWO (Research Foundation Flanders), and the EWI ministry for supporting research in the CLiPS Computational Linguistics Group on computational stylometry. The research described in Sect 3 is sponsored by IWT (Flemish Agency for Innovation by Science and Technology) and the University of Antwerp research fund. Some of the research described explicitly in this paper was done in cooperation with (former) CLiPS colleagues Kim Luyckx, Mike Kestemont, Vincent Van Asch, and Claudia Peersman. The possible errors in interpretation of the results are my own. I am grateful to all CLiPS members for providing an intellectually stimulating research environment.

References

1. van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., Neijt, A.: New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics* 12(1) (2005) 65–77.
2. Stamatatos, E.: A survey of modern authorship attribution methods. *JASIST* 60(3) (2009) 538–556.
3. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. *JASIST* 60(1) (2008) 9–26.
4. Juola, P.: Author attribution. *Foundations and Trends in Information Retrieval* 1(3) (2008) 233–334.
5. Pennebaker, J.: *The Secret Life of Pronouns.* New York: Bloomsbury Press (2011).
6. Fan, J., Kalyanpur, A., Gondek, D., Ferrucci, D.: Automatic knowledge extraction from documents. *IBM Journal of Research and Development* 56(3/4) (2012) 1–10.
7. Liu, B.: *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers (2012) 180 pages.
8. Sebastiani, F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1) (2002) 1–47.
9. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. *Language Resources and Evaluation* 45 (2011) 83–94.
10. Daelemans, W., Van den Bosch, A.: *Memory-based language processing.* Cambridge: Cambridge University Press (2005).
11. Argamon, S.: Interpreting Burrow's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing* 23(3) (2008) 131–147.
12. Koppel, M., Schler, J., Bonchel-Dokov, E.: Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research* 8 (2007) 1261–1276.
13. Rudman, J.: The state of authorship attribution studies: some problems and solutions. *Computers and the humanities* 31(4) (1997) 351–365.

14. Rudman, J.: The satet of non-traditional authorship studies 2010: some problems and solutions. *Proceedings of the Digital Humanities* (2010) 217–219.
15. Stamou, C.: Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, 23(2) (2008) 181–199
16. Brennan, M., Afroz, S., Greenstadt, R.: Adversarial Stylometry: circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security* 15(3) (2012) 12:1–22
17. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1) (2011) 35–55.
18. Grieve, J.: Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22(3) (2007) 251–270.
19. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. *Proceedings 21st international conference on Machine Learning* (2004) 489–495.
20. Koppel, M., Schler, J., Argamon, S., Winter, Y. The Fundamental Problem of Authorship Attribution. *English Studies*, 93(3) (2012) 284–291.
21. Luyckx, K.: *Scalability Issues in Authorship Attribution.* Antwerp: UPA (2010).
22. Daumé III, H., Marcu. D.: Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research* 26 (2006) 101–126.
23. Kestemont, M., Luyckx, K., Daelemans, W., Crombez, T.: Cross-genre authorship verification using unmasking. *English Studies* 93(3) (2012) 340–356.
24. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic plagiarism analysis. *Language Resources and Evaluation* 45(1) (2011) 63–82.
25. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation. *Proceedings of the 2006 EMNLP* (2006) 482–491.
26. Koppel, M., Argamon, S., Shimoni, S.: Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4), (2003) 401–412.
27. Peersman, C., Daelemans W., Van Vaerenbergh L.: Predicting Age and Gender in Online Social Networks. *3rd International Workshop on Search and Mining User-generated Contents (SMUC2011)* (2012) 37–44.
28. Peersman, C., Vaassen F., Van Asch V., Daelemans W.: Conversation Level Constraints on Pedophile Detection in Chat Rooms. *CLEF 2012 Conference and Labs of the Evaluation Forum* (2012) 1–13.
29. Luyckx, K., Vaassen F., Peersman C., Daelemans W. Fine-Grained Emotion Detection in Suicide Notes: A Thresholding Approach to Multi-Label Classification. *Biomedical Informatics Insights* 2012:5(Suppl. 1) (2012) 61–69.