Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7145–7151
Marseille, 11–16 May 2020

Profiling–UD: a Tool for Linguistic Profiling of Texts

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab
Via G. Moruzzi, 1, Pisa, Italy
{name.surname}@ilc.cnr.it

Abstract

In this paper, we introduce Profiling–UD, a new text analysis tool inspired to the principles of linguistic profiling that can support language variation research from different perspectives. It allows the extraction of more than 130 features, spanning across different levels of linguistic description. Beyond the large number of features that can be monitored, a main novelty of Profiling–UD is that it has been specifically devised to be multilingual since it is based on the Universal Dependencies framework. In the second part of the paper, we demonstrate the effectiveness of these features in a number of theoretical and applicative studies in which they were successfully used for text and author profiling.

Keywords: Linguistic Profiling, Universal Dependencies, Computational Language Variation Analysis

1. Introduction

The availability of large-scale corpora representative of real language use, together with the development of sophisticated Natural Language Processing (NLP) pipelines that make explicit the linguistic structure underlying a text, has encouraged a paradigmatic shift towards a more extensive use of data–driven methods to tackle traditional topics of linguistic research, going from the study of language variation across textual genres, registers as well styles associated with particular authors or historical periods, to the investigation of linguistic complexity from a user–dependent perspective. They have also favoured the development of language technologies focusing on modeling the 'form', rather than the content, of texts.

By relying on different levels of linguistic annotation, it is possible to extract a large number of features modeling lexical, grammatical and semantic phenomena that, all together, contribute to characterize language variation within and across texts. These are the prerequisites of linguistic profiling, a methodology in which – in van Halteren's words – "the occurrences in a text are counted of a large number of linguistic features, either individual items or combinations of items. These counts are then normalized [...]" in order to detect and quantify differences and similarities across texts representative of distinct language varieties (van Halteren, 2004). Following this approach, the linguistic structure of a text is analyzed to extract relevant features, and a representation of the text is constructed out of occurrence statistics of these features, be they absolute or relative frequencies or more complex statistics. In linguistic profiling, each text or collection of texts is thus assigned a feature–based representation covering different levels of linguistic description.

Nowadays, linguistic profiling as described above lies at the heart of current research in different areas, which share the purpose of reconstructing the linguistic profile underlying linguistic productions originating in specific contexts, e.g. in socio–culturally defined demographic groups or individual authors (also across time), or arising in different situational contexts and meeting distinct communicative goals. Among them, it is worth reporting here:

- *Computational Register Analysis* (Argamon, 2019), which looks at register and genre variation as essentially functional in nature, deriving from the fact that different communicative contexts require different linguistic resources, resulting in (statistically) different language varieties (covering diaphasic but also diamesic variation);
- *Computational Sociolinguistics* (Nguyen et al., 2016), an emerging research line combining paradigms and methods of computational linguistics and sociolinguistics and focusing on the social dimension of language and the variation inherently connected with it (diastratic variation);
- *Computational Stylometry*, aimed at extracting knowledge from texts with the goal of verifying and attributing authorship (Daelemans, 2013);
- measures and models of *natural language complexity*, which are progressively attracting the interest of the Computational Linguistics community, also due to their impact in applicative scenarios addressing the needs of specific classes of users (see, for instance, the automatic readability assessment methods and techniques surveyed in Collins-Thompson (2014)).

Despite the different ultimate goals, research in the above mentioned areas shares the need to extract – to put it in Daelemans' terms (Daelemans, 2013) – meta-knowledge from texts, namely what are the features and how they combine together within a specific language variety as opposed to another one of the same nature, be it determined on the basis of the communicative purposes in a given situational context, or of the speaker socio-demographic traits, or of the author, or of the addressee. Meta-knowledge extraction thus consists in associating the feature-based representation of a (collection of) text(s) with a functional context, or with a class of speakers and/or addressees, or with individual authors.

Since the beginning, simple and effective sets of features for register and stylistic text analysis were represented by the relative frequencies of function words taken as indicative of different grammatical choices, or of character n-grams assumed to capture linguistic variation in lexical, grammatical, and orthographic preferences. Both feature types are easy to extract: if the former requires language-specific lists of a few hundred words (including pronouns, prepositions, auxiliary and modal verbs, conjunctions, and determiners), the latter – while lacking explicit linguistic motivation – is language-independent. More recently, significant advances in knowledge extraction from text have been made possible thanks to the development of robust and fairly accurate text analysis pipelines for several languages. This also holds for all the aforementioned scenarios, where NLP-based tools that allow to automatize the process of feature extraction play an important role.

Different packages exist today for register, stylistic or linguistic complexity analysis, making use of different types of features. Among these, the *Stylo* package (Eder et al., 2016) offers a rich and user-friendly suite of functionalities for stylometric analyses. *Stylo* focuses on shallow text features that can be automatically extracted without having to resort to language-dependent annotation tools, namely n-grams at token and character levels. Note, however, that it can also accommodate the output of linguistic annotation tools. Several tools are also available for the assessment of text complexity. A well-known one is *Coh-Metrix*, which computes over 200 measures of cohesion, language, and readability from an input text (Graesser et al., 2014), based on features extracted from multi–level linguistic annotation. In a similar vein, both *TAASSC* (Kyle, 2016) and *L2 Syntactical Complexity Analyzer (L2SCA)*(Lu, 2010) allow computing a number of linguistically–motivated indices related to grammatical complexity at phrasal and sentence levels, which have been proven particularly relevant in studies on first and second language acquisition. While all these tools are conceived for the English language, a notable exception is *SweGram*, a system specifically devised to profile texts in the Swedish language (Näsman et al., 2017).

From this sketchy outline, it emerges that language-independent tools such as *Stylo* make typically use of shallow features which do not need language-specific pre-processing, whereas tools based on a wide range of multi-level linguistic features are typically monolingual. In this paper we present Profiling–UD, a new text analysis tool inspired to the principles of linguistic profiling that can support language variation research by allowing the extraction of more than 130 features, spanning across different levels of linguistic annotation, and modeling phenomena related to the 'form' of a text. Differently from other existing tools, it has been specifically devised to be multilingual since it is based on the Universal Dependencies (UD) representation (Nivre, 2015). In this way, linguistic profiling can be carried out in parallel for different languages and, thanks to the shared annotation scheme, results achieved in different areas can also be analysed cross-linguistically.

The paper is organized as follows. In section 2., the linguistic profiling tool is described with a specific view to the wide typology of features that can be monitored. Section 3. reports the results of different case studies, with the final aim of demonstrating the effectiveness of these features in both theoretical studies and applicative scenarios in which they were successfully used for text and author profiling.

2. Profiling–UD

Profiling–UD[1] is a web-based application inspired to the methodology initially presented in Montemagni (2013) and successfully tested in different case studies (some of which are reported in section 3.), that performs linguistic profiling of a text, or a large collection of texts, for multiple languages.

The tool implements a two–stage process: linguistic annotation and linguistic profiling. The first step, linguistic annotation, is automatically carried out by UDPipe (Straka et al., 2016), a state–of–the–art pipeline available for nearly all languages included in the Universal Dependencies (UD) initiative, which carries out basic pre-processing steps, i.e. sentence splitting and tokenization, POS tagging, lemmatization and dependency parsing. In the second step, a set of about 130 features representative of the linguistic structure underlying the text are extracted from the output of the different levels of linguistic annotation. These features capture a wide number of linguistic phenomena ranging from superficial, morpho–syntactic and syntactic properties, which were proven to be effective in several scenarios focusing on the 'form' of a text. A subset of these features is described in Section 2.1..

In the web–based interface, the user is given the option of either uploading a plain text file (or a collection of files as a zipped folder) or copying the text for the analysis. Before running the analysis, it is required to specify the language of the input text. As mentioned before, the annotation of the text(s) is performed by UDPipe using the available UD model(s), version 2.4, for the input language [2]. When more than one model is available for a given language, the one trained on the biggest available treebank is automatically loaded. For each uploaded text, the result of the annotation stage is a file in the CoNLLU–tab–separated format.

The automatically annotated text(s) are used as input to the further step, performed by the linguistic profiling component, which is based on a set of scripts written in Python defining the rules to extract and quantify the formal properties, a selection of which is described in Section 2.1. below. The output of the linguistic profiling step is represented as a table which, for a given text or collections of texts, associates to each monitored feature the corresponding value. The result is reported in a downloadable file in csv format, with each monitored feature in a separate column.

2.1. Linguistic Features

The set of linguistic features monitored by Profiling–UD are extracted from the different levels of annotation and capture a wide number of linguistic phenomena, which can be grouped as follows:

1. Raw Text Properties
2. Lexical Variety

[1] The tool is available at the following link: http://linguistic-profiling.italianlp.it

[2] https://ufal.mff.cuni.cz/udpipe/models

3. Morphosyntactic information
4. Verbal Predicate Structure
5. Global and Local Parse Tree Structures
6. Syntactic relations
7. Use of Subordination

In what follows, the list of features in each category is reported together with a description of how they are extracted from UD representations and quantified. To exemplify some of them, we will refer to the following sentence, whose output is reported in Figure 1.:
(1) *President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington area.*

1. Raw Text Properties

- *Document length*: length of the document calculated both in terms of the total number of tokens and of the total number of sentences it is constituted of.
- *Sentence length*: average length of sentences in a text or collection of texts, calculated as the average number of tokens per sentence.
- *Word length*: calculated as the average number of characters per word (excluding punctuation).

Sentence length and word length are typically seen as proxies of syntactic complexity and lexical complexity respectively, as testified by traditional formulas developed for the automatic assessment of text readability.

2. Lexical Variety

- A standard metric to assess the lexical variety of a text is constituted by the Type/Token Ratio (TTR), which refers to the ratio between the number of lexical types and the number of tokens within a text. Due to its sensitivity to sample size, this feature is computed for text samples of equivalent length: Profiling–UD computes TTR for both the first 100 and 200 tokens of a text.

3. Morpho–syntactic information

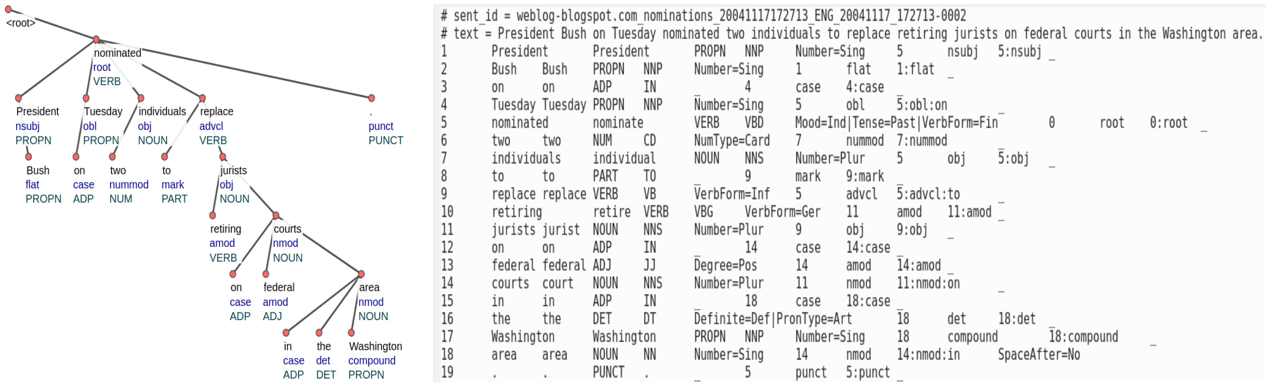
- *Distribution of grammatical categories*: Profiling–UD computes the percentage distribution in the text(s) of the 17 core part-of-speech categories defined in the Universal POS tagset, which are internally subdivided into open class words (i.e. adjective, adverb, interjection, noun, proper noun, verb), closed class words (adposition, auxiliary, coordinating conjunction, determiner, numeral, particle, pronoun, subordinating conjunction), and the class of ‘other’ which includes punctuation and symbols.
- *Lexical density*: this feature refers to the ratio of content words (verbs, nouns, adjectives and adverbs) over the total number of words in a text;
- *Inflectional morphology*: this feature is calculated for lexical verbs and auxiliaries, by taking into account the distribution, for each verb, of the following subset of inflectional UD features, namely *Mood, Number, Person, Tense and Verb(al)Form*.

4. Verbal Predicate Structure

- *Distribution of verbal heads*: calculated as the average number of verbal heads in a sentence, corresponding to the number of propositions co-occurring in it, be they main or subordinate clauses;
- *Distribution of verbal roots*: calculated as the percentage of verbal roots out of the total of sentence roots;
- *Verb Arity*: calculated as average number of instantiated dependency links (covering both arguments and modifiers) sharing the same verbal head, with exclusion of punct(uation) and cop(ula) UD dependencies. Information about the average score is complemented with the distribution of verbal predicates by arity (e.g.verbs with arity=2 are verbs heading 2 dependency links, be they core or non-core arguments in UD parlance). For example, in sentence (1) the average arity score is 2, since the main verb ‘nominated’ has four dependents (‘President’, ‘Tuesday’, ‘individuals’, ’replace’), the first embedded verb ‘replace’ has two (‘to’, ‘jurists’) and the gerund ‘retiring’ has no dependents.

5. Global and Local Parsed Tree Structures

- *Average depth of the syntactic tree*: it corresponds to the mean of the maximum depths extracted for each sentence in a text. The maximum depth of a sentence is calculated as the longest path (in terms of occurring dependency links) from the root of the dependency tree to some leaf. In sentence (1), this feature is equal to 5, corresponding to the five intermediate dependency links that are crossed in the path going from the root of the sentence (‘nominated’) to each of the equidistant leaf nodes, represented by the words ‘in’, ‘the’ and ‘Washington’.
- *Average clause length*: calculated in terms of the average number of tokens per clause, where the number of clauses corresponds to the ratio between the number of tokens in a sentence and the number of either verbal or copular heads.
- *Length of dependency links*: the length of a dependency link is calculated as the number of words occurring linearly between the syntactic head and its dependent (excluding punctuation dependencies). The value associated with this feature corresponds to the average value extracted for all dependencies in a text. This information is complemented with the feature *Maximum dependency link* corresponding to the average length of the longest dependency link for each



```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# text = President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington area.
1       President       President       PROPN   NNP     Number=Sing     5       nsubj   5:nsubj _
2       Bush    Bush    PROPN   NNP     Number=Sing     1       flat    1:flat  _
3       on      on      ADP     IN      _       4       case    4:case  _
4       Tuesday Tuesday PROPN   NNP     Number=Sing     5       obl     5:obl:on        _
5       nominated       nominate        VERB    VBD     Mood=Ind|Tense=Past|VerbForm=Fin        0       root    0:root  _
6       two     two     NUM     CD      NumType=Card    7       nummod  7:nummod        _
7       individuals     individual      NOUN    NNS     Number=Plur     5       obj     5:obj   _
8       to      to      PART    TO      _       9       mark    9:mark  _
9       replace replace VERB    VB      VerbForm=Inf    5       advcl   5:advcl:to      _
10      retiring        retire  VERB    VBG     VerbForm=Ger    11      amod    11:amod _
11      jurists jurist  NOUN    NNS     Number=Plur     9       obj     9:obj   _
12      on      on      ADP     IN      _       14      case    14:case _
13      federal federal ADJ     JJ      Degree=Pos      14      amod    14:amod _
14      courts  court   NOUN    NNS     Number=Plur     11      nmod    11:nmod:on      _
15      in      in      ADP     IN      _       18      case    18:case _
16      the     the     DET     DT      Definite=Def|PronType=Art       18      det     18:det  _
17      Washington      Washington      PROPN   NNP     Number=Sing     18      compound        18:compound     _
18      area    area    NOUN    NN      Number=Sing     14      nmod    14:nmod:in      SpaceAfter=No
19      .       .       PUNCT   .       _       5       punct   5:punct _
```

(a) Graphical visualization.

(b) CoNLL-U format.

Figure 1: Linguistic annotation of example sentence (1).

sentence in a given text. To give an example, in Sentence (1) there are 17 dependency links[3]. Eight links have a one–token distance: [‘**President**’,‘Bush’], [‘on’,‘**Tuesday**’], [‘Tuesday’,‘**nominated**’], [‘two’,‘**individuals**’], [‘to’,‘**replace**’], [‘retiring’,‘**jurists**’], [‘federal’,‘**courts**’], [‘Washington’,‘**area**’]. Four links have a two–token distance: [‘**nominated**’,‘individuals’], [‘**replace**’,‘jurists’], [‘on’,‘**courts**’], [‘the’,‘**area**’]. Two links have a three–token distance: [‘**jurists**’,‘courts’], [‘in’,‘**area**’]. Three links show the maximum four–token distance: [‘President’,‘**nominated**’], [‘**nominated**’,‘replace’], [‘**courts**’,‘area’]. The average value, calculated as the ratio between the sum of all distances over the total number of links, is 2.

- *Average depth of embedded complement chains governed by a nominal head*: this feature refers to the depth of embedded complement ‘chains’ sharing the same nominal head and including either prepositional complements or nominal and adjectival modifiers. Its value corresponds to the average depth of complex nominal chains extracted from all sentences in a given text. In Sentence (1), the depth of the nominal chain headed by the noun ‘jurists’ is equal to 2; as it can observed in the graphical visualization in Figure 1(a), the chain covers two embedded prepositional modifiers (‘on federal courts’ and ‘in the Washington area’), both governed by the noun ‘jurists’. This information is complemented by additional features concerning the distribution of complex nominal constructions by depth.

- *Word order phenomena*: in Profiling-UD monitored word order phenomena are circumscribed to the main elements of the sentence, i.e. the subject and the object. Under this heading, there are features corresponding to the relative order of subject or object with respect to the verb, with the associated probability distribution. This feature is expected to capture word order variation across languages and – within the same language – across varieties of language use. Consider, for example, the distribution of pre– and post–verbal nominal subjects and objects across a selection of UD treebanks representative of three language families, i.e. Germanic, Romance and Slavic. As it can be seen in Figure 2, with the only exception of Afrikaans, all considered languages tend to prefer patterns compliant with the SVO order. However, while Germanic and Romance languages display a lower percentage of nominal subjects following the verb (17.36% and 17.63% respectively), the post verbal position is more frequent in the Slavic treebanks (30.05%). Fewer differences are reported for what concerns the position of the object with respect to the verb, even if among the Germanic treebanks Afrikaans shows an opposite trend characterized by a clear preference for pre-verbal objects (almost 80%).

6. Syntactic Relations

- *Distribution of dependency relations*: this feature refers to the percentage distribution of the 37 universal relations in the UD dependency annotation scheme.

7. Subordination phenomena

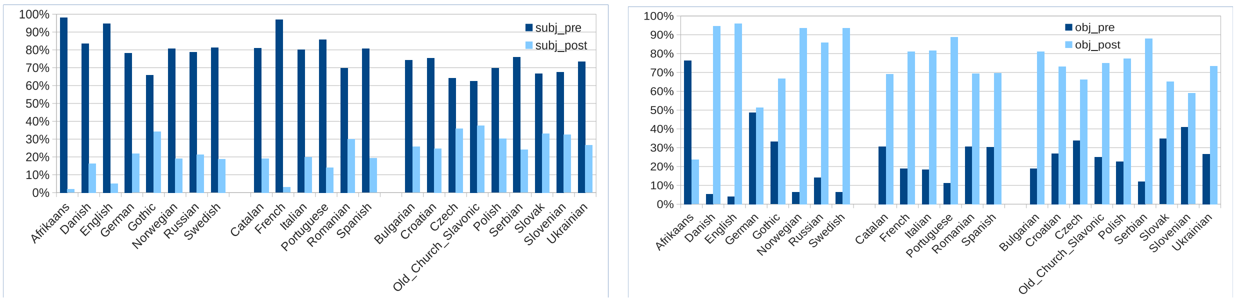
- *Distribution of subordinate and main clauses*: this feature is calculated as the percentage distribution of main vs subordinate clauses as defined in the UD scheme [4]. The values can be combined to calculate the ratio between the two.

- *Relative order of subordinates with respect to the verbal head*: as for subjects and objects, this feature is calculated as the percentage distribution of subordinate clauses in post–verbal and in pre–verbal position.

- *Average depth of embedded subordinate clauses*: given the subordinate clause sub–tree, a subordinate ‘chain’ is calculated as the number of subordinate clauses recursively embedded in the top subordinate clause. In addition to the average value of the chain depth, the percentage distribution of subordinate chains by depth is also provided.

[3] In all examples, the head is highlighted in bold.

[4] https://universaldependencies.org/u/overview/complex-syntax.html#subordination



(a) Pre and post verbal nominal subjects.

(b) Pre and post verbal objects.

Figure 2: Distribution of pre and post nominal subjects and objects across UD treebanks of Germanic (left group), Slavic (group in the middle) and Romance (rigth group) languages.

Sentence (1) is articulated into a main clause and a subordinate clause governed by the verbal root 'nominated'; the adverbial subordinate clause headed by 'replace' occurs in post–verbal position and does not contain in its turn embedded subordinates.

3. Case Studies

In order to exemplify the potential of Profiling–UD and in particular the reliability and effectiveness of the information extracted with it, we summarise below the results achieved in different application scenarios in which the proposed linguistic profiling methodology has been successfully applied. The case studies proposed below, mostly carried out with respect to Italian, show how the wide set of linguistic features extracted by Profiling–UD can be successfully exploited in a variety of text and author profiling tasks, covering different aspects of register, sociolinguistic, stylometric and complexity analysis.

3.1. Text Profiling

Text Readability. This task falls in the area focusing on the analysis of language complexity. In recent years, multi-level linguistic profiling of texts started being progressively used to assess the degree of reading difficulty faced by human readers with different backgrounds. According to the wide literature on text readability, the possibility of modeling linguistic phenomena capturing different aspects of text difficulty has played a main role for Automatic Readability Assessment (Collins-Thompson, 2014). A case study focused on Italian has shown that the set of linguistic features introduced in Section 2.1. can be usefully exploited to automatically assess the degree of readability of texts (Dell'Orletta et al., 2011). As fully described by the authors, different combinations of features contribute in a different way to the automatic assessment of text readability, at document and sentence levels. As expected, raw text features traditionally considered in the literature as a proxy of lexical and syntactic complexity (Rogers and Chissom, 1973; Lucisano and Piemontese, 1988) resulted to have a discriminating power in distinguishing easy- and difficult-to-read documents and sentences. However, more complex linguistic features, such as lexical, morpho–syntactic and syntactic ones, allow achieving higher automatic classification performances since they are able to account for finer–grained aspects of text complexity.

A further investigation carried out by Dell'Orletta et al. (2014) to identify the most discriminative features in document vs sentence readability assessment revealed that *i)* sentence readability analysis requires a higher number of features, i.e. about twice the ones required for documents, and *ii)* the most predicting features for sentence readability refer to local sentence complexity, e.g. features such as the arity of verbal predicates, or the distribution of preverbal as opposed to post-verbal subjects, or of post-verbal objects as opposed to pre-verbal ones. On the other hand, features capturing more global syntactic phenomena, such as the pre-verbal position of subordinate clauses, are more relevant for what concerns document classification.

Textual Genre. The set of morpho–syntactic and syntactic features monitored by Profiling–UD turned out to play a key role also to predict the textual genre of a document, a task which is part of register analysis as reported above. As discussed by Cimino et al. (2017) in a comparative analysis of four traditional textual genres (i.e. educational material, newspaper articles, literary texts educational material, and scientific papers), these features have a higher discriminative power if compared with the simple information provided by lemma unigrams. Interestingly enough, among them syntactic features turned out to play a key role for the classification of textual genres. However, the typology of features mostly contributing to the recognition of a given genre as opposed to another one can change across genres. For instance, features characterizing the overall sentence structure, i.e. the parse tree depth and the maximum length of dependency links, play a key role in the classification of the Literature and Journalism genres. Other syntactic features which turned out to play a relevant role are concerned with: the relative ordering of subject and object with respect to the verbal head; or the use of passive voice which is highly ranked in the characterization of scientific writing and newspaper articles, and less relevant for the classification of the Literature genre.

3.2. Author Profiling

Since the seminal work by Argamon et al. (2003), it has been shown that it is possible to identify sociolinguistically defined classes of authors based on their use of language.

Detecting writing characteristics shared by authors in order to predict their gender, age, native language, personality, etc. is currently receiving a growing interest in the Computational Linguistics community as it is also testified by the first shared task organized in 2013 on *Author Profiling at PAN 2013* (Rangel et al., 2013).

In this section, we will show how the set of linguistic features described in Section 2.1. has been successfully used as fingerprints of some aspects of writing style characterizing classes of authors.

Author's Gender. Possible differences between female and male writing style have been widely investigated in the literature also with respect to the interference of textual genre (Rangel et al., 2016). The *4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations* focused on English, Spanish and Dutch languages, followed by the *GxG at EVALITA 2018 shared–task: Cross–genre gender prediction* for Italian (Dell'Orletta and Nissim, 2018), offered the opportunity to investigate whether there are gender-specific characteristics that remain constant across different textual genres. In both competitions, it turned out that cross-genre prediction of author's gender is a quite challenging task suggesting that females and males might use a different writing style according to textual genre.

To get a better understanding of linguistic phenomena possibly characterizing male vs female writing, Cocciu et al. (2018) carried out linguistic profiling experiments of Italian data used in the GxG task. The analyses revealed that independently from the textual genre, women tend to write using a more informal style, characterized e.g. by shorter clauses, shorter dependency links and shallower syntactic trees, as well as the prevalent use of subordinate clauses in canonical (i.e. pre-verbal) position. On the contrary, male sentences are longer, syntactically articulated and contain a higher percentage of nouns and proper nouns, denoting a more objective writing style.

Author's Age. Since the *Author Profiling Task at PAN 2013* (Rangel et al., 2013) both the identification of gender and of age have been considered two aspects of the author profiling problem. The linguistic features underlying Profiling–UD resulted to be effective also in this scenario, as shown by Maslennikova et al. (2019). The authors built a corpus of Italian forums on different topics balanced by 6 different age groups (ranging from $\leq$20 years to $\geq$61). The set of experiments they performed was aimed at showing the impact of the different typology of considered features in the classification of forums by user's age. Comparing the performance of the classification models, it turned out that lexical information is the most predictive one[5]. However, sentence structure identified by the set of syntactic and morpho-syntactic features plays also a significant role. Indeed, the classification model relying only on this typology of features outperforms the baseline and for some topics (e.g. sport) achieves the best accuracy.

Author's Native Language. "Identifying an author's native language is a type of authorship attribution problem. Instead of identifying a particular author from among a closed list of suspects, we wish to identify an author class, namely, those authors who share a particular native language" (Koppel et al., 2005). Koppel's definition highlights how profiling authors according to their native language can be seen as a process of detection of fingerprints of groups of authors. Since the organization of the *First Shared Task on Native Language Identification* (Tetreault et al., 2013), stylistic characteristics of L2 writings have been used to model L1 features and predict the native language of the writer of a given document. Also in this scenario, Profiling–UD features turned out to be effective in i) classifying the L1 of the writer and ii) reconstructing the linguistic profile of L1 starting from L2 productions (Cimino et al., 2013). For instance, the authors highlighted that L1s belonging to the same language family (e.g. Japanese and Korean), or contact languages (e.g. Hindi and Telugu), show closer distributions of features. The authors carried out an additional analysis aimed at investigating whether the linguistic information extracted from either the whole document or each single sentence might contribute differently to the task (Cimino et al., 2018). Similarly to what observed for the text readability assessment scenario, it was shown that for document classification purposes low level features, such as words n-grams, are sufficient enough to predict L1, while morpho-syntactic and syntactic features are more effective for sentence classification.

4. Conclusion

We presented Profiling–UD, the first tool for multilingual linguistic profiling based on the Universal Dependency framework. It allows the extraction of a wide set of features acquired from different levels of linguistic annotation. The consistent annotation of similar constructions across languages guaranteed by UD makes the process of features extraction applicable to different languages: ongoing work includes an extension of the extraction rules aimed at handling similar constructions in typologically different languages.

Profiling–UD can be usefully exploited by scholars in the areas of digital humanities and theoretical linguistics, to automatically extract a wide range of sophisticated linguistic features on the basis of which to carry out large–scale investigations on language variation from different perspectives, ranging from register and sociolinguistic studies to stylistic or linguistic complexity analyses. Since UD is a framework featuring consistent linguistic annotation across different languages, Profiling–UD creates the prerequisites for cross–lingual studies focusing on the 'form' rather than the content of texts. The linguistic profile automatically reconstructed by Profiling–UD can also be exploited for a variety of text and author automatic classification tasks within different application scenarios. Last but not least, the linguistic knowledge encoded by Profiling–UD can also support computer scientists who may benefit from an explicit representation of linguistic phenomena useful to develop a wide range of language–related tasks, as well to address open issues related to the interpretability of neural networks models.

[5] For the purpose of these experiments the author considered word and lemma n-grams as lexical features.

5. Bibliographical References

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3).

Argamon, S. (2019). Computational register analysis and synthesis. *Register Studies*, 1(1):100–135.

Cimino, A., Dell'Orletta, F., Venturi, G., and S., M. (2013). Linguistic profiling based on generalâpurpose features and native language identification. In *Proceedings of Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–215, Atlanta, Georgia, June.

Cimino, A., Wieling, M., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2017). Identifying predictive features for textual genre classification: the key role of syntax. In *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Rome, December.

Cimino, A., Dell'Orletta, F., Brunato, D., and Venturi, G. (2018). Sentences and documents in native language identification. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Turin, December.

Cocciu, E., Brunato, D., Venturi, G., and Dell'Orletta, F. (2018). Gender and genre linguistic profiling: a case study on female and male journalistic and diary prose. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Turin, December.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(1):97–135.

Daelemans, W. (2013). Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, pages 451–462, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dell'Orletta, F. and Nissim, M. (2018). Overview of the evalita 2018 cross-genre gender prediction (gxg) task. In *Proceedings of EVALITA '18, Evaluation of NLP and Speech Tools for Italian*, December, date =.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). READ-IT: assessing readability of italian texts with a view to text simplification. In *Proceedings of Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edimburgo, UK, July.

Dell'Orletta, F., Wieling, M., Montemagni, S., Venturi, G., Cimino, A., and Montemagni, S. (2014). Assessing the readability of sentences: Which corpora and features? In *Proceedings of Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 163–173, Baltimore, Maryland, USA, june.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with r: A package for computational text analysis. *The R Journal*, 8(1):107–121.

Graesser, A., McNamara, D., Cai, Z., Conley, M., Li, H., and Pennebaker, J. (2014). Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 2:210–229.

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous author's native language. *Intelligence and Security Informatics*, 3495:209–217.

Kyle, K. (2016). Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. In *Doctoral Dissertation.*

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15 (4):474–496.

Lucisano, P. and Piemontese, M. E. (1988). Una formula per la predizione della difficolta dei testi in lingua italiana. *Scuola e Città*, 3:57–68.

Maslennikova, A., Labruna, P., Cimino, A., and Dell'Orletta, F. (2019). Quanti anni hai? age identification for italian. In *Proceedings of 6th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Bari, November.

Montemagni, S. (2013). Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. pages 145–172.

Näsman, J., Megyesi, B., and Palmér, A. (2017). Swegram - a webbased tool for automatic annotation and analysis of swedish texts. In *Proceedings of 21st Nordic Conference on Computational Linguistics (Nodalida).*

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: a survey. *Computational Linguistics*, 42:537–593.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh, editor, *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G. (2013). Overview of the author profiling task at PAN 2013. In *CLEF 2013 Evaluation Labs and Workshop â Working Notes Papers*, pages 1–13, September.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Pottast, M., and Stein, B. (2016). Overview of the 4th author profiling task at pan 2016: Cross-genre evaluations. In *CLEF 2016 Labs Labs and Workshop â Working Notes Papers*, pages 750–784, Evora, Portugal, September.

Rogers, L. and Chissom, B. S. (1973). Derivation of new readability formulas for navy enlisted personnel. Research branch report, Chief of Naval Training, Millington, TN.

Straka, M., Hajic, J., and Strakova, J. (2016). UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC).*

Tetreault, J., Blanchard, D., and Cahill, A. (2013). Summary report on the first shared task on native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, GA, USA, June.

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics*, pages 200–207.