Building a Corpus for Personality-dependent Natural Language Understanding and Generation

R.M.S. Ramos, G.B.S. Neto, B.B.C. Silva, D.S. Monteiro, I. Paraboni, R.F.S. Dias
University of São Paulo, School of Arts, Sciences and Humanities
São Paulo, Brazil
{ricellimsilva,georges.stavracas,barbarab.claudino,danisamon,ivandre.paraboni,rafaelsandronidias}@gmail.com

Abstract

The computational treatment of human personality - both for the recognition of personality traits from text and for the generation of text so as to reflect a particular set of traits - is central to the development of NLP applications. As a means to provide a basic resource for studies of this kind, this article describes the b5 corpus, a collection of controlled and free (non-topic specific) texts produced in different (e.g., referential or descriptive) communicative tasks, and accompanied by inventories of personality of their authors and additional demographics. The present discussion is mainly focused on the various corpus components and on the data collection task itself, but preliminary results of personality recognition from text are presented in order to illustrate how the corpus data may be reused. The b5 corpus aims to provide support for a wide range of NLP studies based on personality information and it is, to the best of our knowledge, the largest resource of this kind to be made available for research purposes in the Brazilian Portuguese language.

Keywords: Corpora, Personality, Big Five

1. Introduction

In recent years, the development of so-called intelligent systems has devoted a great deal of attention to the computational treatment of human personality. This interest may be explained, among other reasons, by the practical observation that users of computer systems not only attribute human traits to the systems they interact with, but they also prefer those systems that present traits similar to their own (Mairesse et al., 2007).

Fundamental personality traits are consistently reflected in the language choices made by individuals when communicating. For instance, an individual with narcissistic traits might make frequent use of first-person expressions ('I', 'for me', etc.). The relation between personality and natural language is the focus of a large body of work in the Psychology field, and it is perhaps best summarised by the *Big Five* personality factors (Goldberg, 1990) - Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism - which are widely accepted as an adequate basis for the representation of human personality.

Given its linguistic motivation, the *Big Five* model provides a theoretical basis for the computational treatment of personality on at least two fronts: the automatic recognition of personality traits from text (which is a language understanding task), and the generation of text in order to reproduce certain personality traits of interest (which is a natural language generation (NLG) task). Knowing the personality traits of an individual (e.g., from his/her social network status updates) has many obvious applications, including staff recruitment, credit analysis etc. In addition to that, personality information may also guide the automatic generation of personalised content, the modelling of psychologically plausible virtual agents (e.g., intelligent tutors, game characters, etc.) and human-computer dialogue applications in which a high degree of realism and engagement is required.

Personality-oriented language understanding and generation are considerable research challenges and, despite their complementary nature (for example, in applications of human-computer dialogue), will usually have a common starting point: a basic resource from which we may establish mappings from linguistic features to personality traits. Based on this observation, this article presents the *b5* corpus of texts produced in multiple communicative tasks and accompanied by inventories of personality of their respective authors. The corpus is, to the best of our knowledge, the largest resource of this kind available for the Brazilian Portuguese language, and it intends to provide support for a wide range of NLP studies based on personality traits.

The rest of this paper is structured as follows. After a brief background discussion (Section 2), the work focuses on the corpus collection task (Section 3) and on its various components (Section 4). Preliminary results of personality recognition from the corpus text are presented for illustration purposes (Section 5). This is followed by a discussion on possible applications and extensions of the present work (Section 6).

2. Background

The *Big Five* model (Goldberg, 1990) comprises five fundamental dimensions of the human personality - Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism - that may be estimated by using a wide range of methods, the most common being the use of personality inventories. Among many inventories developed for the *Big Five* model, the need for a fast assessment tool led to the proposal of the *BFI* inventory (John et al., 1991).

The BFI inventory has been replicated in dozens of other languages, including some studies dedicated to our target language, Brazilian Portuguese. In particular, the study in (de Andrade, 2008) validated the *BFI* for Brazilian Portuguese through factorial analysis of a sample of 5,089 respondents from all regions of the country. The inventory considered in (de Andrade, 2008) will be the basis of the present work as well.

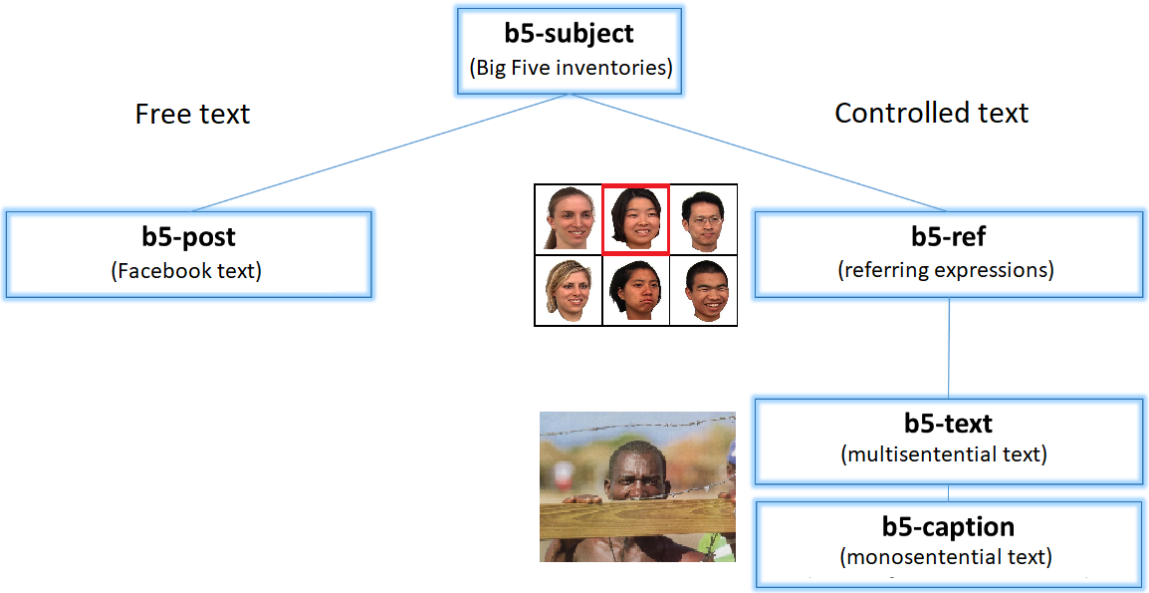

Figure 1: The b5 corpus structure.

The information provided by the *BFI* enables the investigation of a range of issues related to the computational treatment of human personality. A detailed discussion of these applications would be beyond the scope of this paper, but includes the recognition of personality traits from text on social networks (Iacobelli et al., 2011; Celli, 2012; Álvarez-Carmona et al., 2015) and the generation of text based on a target personality (Mairesse and Walker, 2011).

Applications of this kind will usually rely on text corpora annotated with personality information. An example of resource of this kind is *myPersonality*, a large database of Facebook status updates for the English language and corresponding Big Five information about their authors. We are not aware, however, of any similar resources for our target language (Brazilian Portuguese).

In addition to the lack of language resources in this target language, we notice that existing resources are usually devoted to personality recognition applications, but they may be less suitable for personality-dependent language generation (e.g., (Mairesse and Walker, 2011)), in which case it may be necessary to have access not only to the text produced by different individuals (i.e., with different personality traits) but also to the context within which the text was produced. These observations lead us to collect a novel resource for personality-based Portuguese language generation and understanding, hereby called the *b5* corpus.

3. Corpus structure

The *b5* corpus is a dataset containing texts and self-reported personality inventories of their authors. This consists of an author's knowledge base called *b5-subject* and four text databases (or subcorpora) called *b5-post*, *b5-ref*, *b5-text*, and *b5-caption* discussed below. An overview of this organisation is illustrated in Figure 1.

The personality inventory that accompanies the collected texts is a Brazilian Portuguese version of the 44-item *BFI* developed for the English language (John et al., 1991), and presented in (de Andrade, 2008). The set of inventories and additional participant's demographics are represented as the corpus *b5-subject* knowledge base.

From the set of inventories, we computed the five basic factors and, as proposed in (Soto and John, 2009), two additional facets each: (Extraversion) Assertiveness and Activity, (Agreeableness) Altruism and Compliance, (Conscientiousness) Order and Self-discipline, (Neuroticism) Anxiety and Depression, and (Openness to experience) Aesthetics and Ideas.

As shown in Figure 1, the *b5* corpus conveys four text categories divided into two general classes: *free* text obtained from Facebook status updates of each participant, and three types of *controlled* text obtained from a series of in-person data collection tasks. The collection of both free and controlled text is motivated by the dual purpose of the corpus, that is, by our long-term goal of reusing the data both in language understanding and language generation studies. The use of free text is mainly motivated by the specific needs of certain types of application, such as the recognition of personality from text on social networks, whereas the use of controlled text is required for a range of Natural Language Generation (NLG) studies.

The free text dataset constitutes the *b5-post* subcorpus. Controlled texts constitute the *b5-ref* subcorpus of referring expressions, and the *b5-text* and *b5-caption* subcorpora of multi- and mono-sentential descriptions. Since not all participants of the data collection completed every task, each subcorpus may include text produced by a different subset of individuals.

Personality inventories, free and controlled text were collected through a Facebook application and/or in-person experiments. The Facebook application allowed users to respond the personality inventory and, simultaneously, performed the collection of their status updates (upon consent). For a subset of subjects, an offline version of the inventory was made available and, instead of collecting Face-

book text, this was followed by an in-person experiment to elicit controlled text.

As in (Schwartz et al., 2013) and others, Facebook text comprises our major source of knowledge for investigating the relationships between personality and language use. However, since the *b5* corpus intends to provide support to text generation studies as well, the corpus also includes text produced under controlled conditions, in which case not only the text produced by the human subjects is available, but also the original context from which the text was elicited in the first place.

Following much of the work on NLG, controlled text was elicited from visual stimuli represented by images widely used in Psycholinguistics. In the present work, images were taken from the GAPED (Dan-Glauser and Scherer, 2011), Face Place (Righi et al., 2012) and Greebles (Gauthier and Tarr, 1997) image databases[1]. Based on the selected stimuli, participants were requested to produce text in three natural language production tasks: reference production, multi- and mono-sentential scene description. These tasks are described in more detail in the next section.

Both free and controlled texts were subject to spell-checking and basic pre-processing procedures[2]. These included the removal of hashtag and special characters, and the treatment of compound terms (e.g., 'a-m-a-z-i-n-g' or 'HappyDaysAhead'), among others.

For reasons of anonymity, and also to provide a minimal level of normalisation, the text was also subject to a number of replacement operations. In particular, proper names were replaced by a $NAME$ identifier, numeric expressions were replaced by $NUMBER$ and laugh and emotion expressions (e.g., 'yay', 'ouch', 'haha', 'LoL' etc.) were replaced by $LAUGH$, (negative) $EMOTION-$, (positive) $EMOTION+$ or (ambivalent) $EMOTION*$. Some of these operations are however only relevant in the case of free *b5-post* text since expressions of this kind did not generally occur in the controlled texts.

Table 1 presents the number of subjects, sentences (or status updates, in the case of the *post* subcorpus), items (words, punctuation symbols, etc.), and word types in the corpus.

Subcorpus	Subjects	Sentences	Items	Types
post	1,019	194,382	2,219,585	866,243
text	151	1,510	84,463	37,210
caption	151	1,510	4,896	4,121
ref	152	4,558	64,518	18,700

Table 1: Textual data in the *b5* corpus.

4. The b5 components

The corpus consists of four text subcorpora - *b5-post*, *b5-ref*, *b5-text* and *b5-caption* - labelled with their corresponding author's identifiers. Using these identifiers, it is possible to retrieve author's personality scores and additional demographics from the *b5-subject* knowledge base.

b5-subject contains 1082 personality inventories and partial author information regarding gender, age, background, degree of religiosity (on a 1-5 scale) and undergraduate course information. Gender information is known for 1081 (99.9 %) subjects, being 597 (55.2 %) female. Age is known for 810 (74.9 %) subjects, ranging from 18 to 61 years (average of 24.6 years).

Figure 2 illustrates personality distribution across corpus participants.

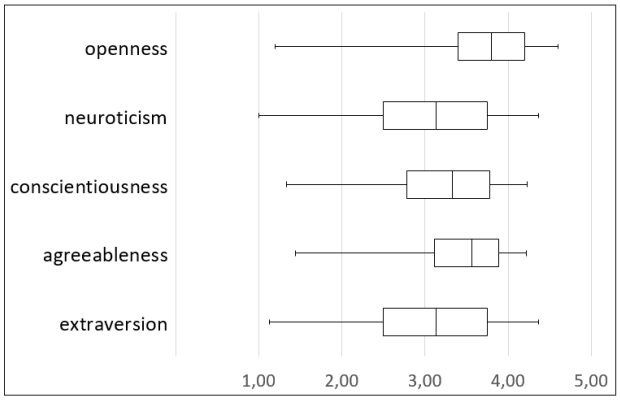

Figure 2: Personality distribution.

Details of the age distribution are presented in Figure 3.

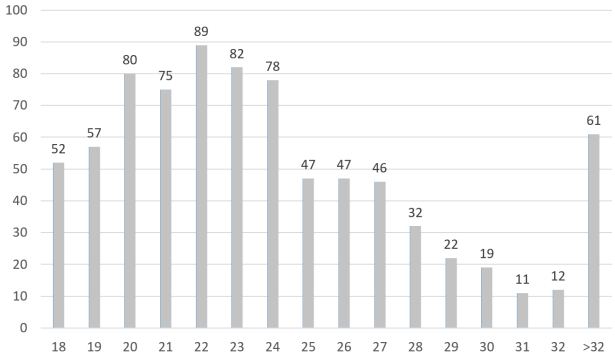

Figure 3: Age distribution.

The individual text components of the *b5* corpus are described in the following sections.

4.1. Facebook status updates *b5-post*

The *b5-post* subcorpus was built for the study and development of computational models of personality recognition and author profiling (e.g., gender or age recognition etc.) from social networks text. The corpus contains Facebook status updates from participants who filled out the personality inventory using the purpose-built application. For each subject, up to 1,000 Facebook status updates were collected. Users with little or no Facebook activity were discarded, resulting in a corpus of 1019 texts.

[1] Face Place and Greeble images are courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University.

[2] Although certain spelling mistakes might be indicative of personality traits, this possible source of knowledge was discarded as a means to enable the use of the corpus in studies of NLG as well, whose focus is usually the generation of correct text.

4.2. Referring Expressions *b5-ref*

The *b5-ref* subcorpus was built for the study of the effects of human personality on the generation of referring expressions (REG), which is an active research topic in NLG (Krahmer and van Deemter, 2012). REG is hereby understood both as the task of determining the semantic contents of definite descriptions (or *what to say* about the intended referent), and as the surface realisation task of these expressions (or *how to say it* in a target language).

As in much of the existing work on data collection for REG (Gatt et al., 2007; Dale and Viethen, 2009; Paraboni et al., 2017a), the *b5-ref* corpus was implemented as a language production task in which subjects were requested to distinguish a certain target from distractor objects in a given context by making use of a definite description. Unlike REG corpora based on simplified domains (e.g., geometric objects), however, *b5-ref* makes use of stimulus images that may arguably make differences across personality traits more explicit. More specifically, the referential contexts under consideration convey images extracted from *Face Place* (Righi et al., 2012), a collection of realistic human photographs annotated with affective and physical attributes.

An example of stimulus image of this kind is illustrated in Figure 4.

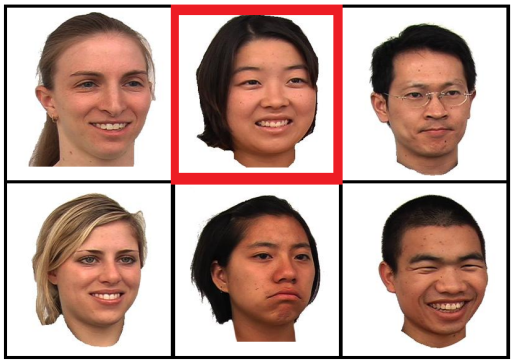

Figure 4: Stimulus image built from *Face Place*.

Given a series of contexts of this kind, subjects were instructed to complete a sentence in the form 'The person / entity highlighted in red is the ... ', which elicited a response in the form of a single referring expression. In the present example, this could be done, for instance, by making use of expressions such as 'the smiling Asian girl' or 'the only girl with dark hair who is smiling', among many other possibilities.

Subjects were instructed to imagine that they were describing each face to a person who could not see their own screen, and that for that reason they should avoid making reference to screen positions (e.g., 'the Asian girl *on the top row*). As a means to reduce the monotonicity of the task, the main stimuli were interleaved with filler images depicting Greebles (Gauthier and Tarr, 1997), since these objects are particularly difficult to identify based on their physical features alone. An example of stimulus image built from Greebles is illustrated in Figure 5.

The *b5-ref* descriptions were produced by 152 subjects, being 86 (56.6 %) female and on average 25.8 years-old (minimum 18 and maximum 59). In its current version, the *b5-ref* corpus contains 1810 Face Place definite descriptions. Given the purpose of the data - for studies on personality-based REG - the *b5-ref* subcorpus differs from the other (purely textual) portions of the *b5* corpus in that these expressions include semantic annotation represented as properties (or attribute-value pairs) as in *gender-female*. To this end, the elicited descriptions were labelled according to a 27-attribute annotation scheme based on the most frequent types of information observed in the corpus.

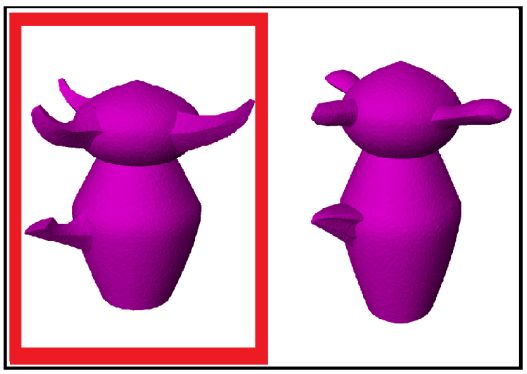

Figure 5: Filler image built from *Greebles*.

Attribute values were partly obtained from *Face Place*, and partially obtained from manual annotation, including both physical (e.g., skin colour, hair length etc.) and affective (e.g., negative and positive emotions) properties. Thus, *b5-ref* is a semantically-annotated REG corpus in the traditional sense, i.e., not unlike TUNA (Gatt et al., 2007) and others, only with additional information about the personality of every speaker.

The annotation scheme disregarded attributes that were not added for the explicit purpose of identification (as in 'the person *who seems to have good taste in clothes*'), and this information was therefore not annotated (although it still remains available from the original text in the corpus). Moreover, as a means to avoid annotating an overly large number of sparse attributes (which may be of little interest from a REG perspective), certain attributes were combined into more general classes according to their semantic affinity. For instance, all references to facial hair (e.g., beard, moustache, goatee etc.) were represented as a single attribute *facial.hair* with possible values *yes / no* indicating simply that there was a general reference to this class of related concepts.

In addition to that, attributes whose value could not be objectively determined (e.g., whether a certain face shape may be considered 'round' or not) were modelled as having only the value *others*. This is intended to represent attributes that have no discriminatory value (e.g., because any of the faces presented as stimulus may be considered, to some extent, as having a round shape), and it carries non-trivial consequences for the design of REG algorithms that favour the selection of discriminatory information and/or pay regard to referential overspecification (Paraboni et al., 2017b).

The relative subjectivity of certain attributes was treated as evenly as possible by considering the information provided by Face Place, if available. Thus, for instance, properties

related to ethnic type (black, Asian etc.) or those that represent emotions (happy, sad etc.) were annotated with their default Face Place values or, if unavailable, according to the judgement of the majority of the 152 participants of the data collection task. Moreover, all descriptions of a given image showing, e.g., a short-haired person (according to the information provided by *Face Place*) were annotated as *hair.length-short* even if a particular individual described it as being long. In other words, the annotated value is meant to model the reference to the *hair.length* attribute, but not necessarily of the actual (*short* or *long*) value chosen by each individual.

Infrequent information was generally omitted from the annotation scheme as well, and it was therefore not recorded. This included references to 'only' (e.g., 'the only smiling Asian girl'), degree modifiers (e.g., 'very', 'slightly' etc.), comparatives (e.g., 'larger than') and references to a second person in the scene (e.g., 'next to a blond girl').

The *b5-ref* subcorpus is provided as two main components: a set of XML files representing the expressions produced by every participant, and the full semantic specification of each of the 12 stimulus scenes. This representation is similar to (Gatt et al., 2007) and many other REG projects. An example is illustrated in Figure 6.

```
<TRIAL ID="7" SPEAKER="31">

 <CONTEXT ID="5">
  <ATTRIBUTE-SET STRING="the smiling asian woman ">
   <ATTRIBUTE NAME="smile" VALUE="yes" />
   <ATTRIBUTE NAME="race" VALUE="asian" />
   <ATTRIBUTE NAME="gender" VALUE="female" />
  </ATTRIBUTE-SET>
 </CONTEXT>

...
<\TRIAL>
```

Figure 6: An annotated referring expression in *b5-ref*.

In the *b5-ref* data, personality traits have been shown to affect both the contents and the surface form of referring expressions. Preliminary results of a machine learning REG model based on *b5-ref* data in (Paraboni et al., 2017c) suggest that the selection of non-discriminatory attributes (e.g., the property of 'being young', which is shared by all objects in the *b5-ref* domain and it is therefore not discriminatory) is particularly influenced by the speaker's personality traits, an effect that is less evident in the case of discriminatory (or more perceptual) attributes. Moreover, results from a personality-dependent lexical choice model built from *b5-ref* in (Lan and Paraboni, 2018) showed that the lexicalisation of the most frequent properties (i.e., those for which there is sufficient data in the corpus) greatly improves when personality information is taken into account.

4.3. Scene descriptions *b5-text* and *b5-caption*

Unlike *b5-ref*, the *b5-text* and *b5-caption* subcorpora are primarily intended for the study of more general issues of personality-based text production, such as document planning, text-to-text generation, and summarisation. To this end, the data collection experiment included two scene description subtasks: a detailed version in the form of multi-sentential text, and a short version in the form of a single sentences similar to picture captions.

The visual stimuli employed in both cases were taken from *GAPED* (Dan-Glauser and Scherer, 2011), a database conveying images classified by valence and normative significance, and designed to arouse different degrees and types of reaction. The image description task made use of 10 GAPED images with valence values selected at regular intervals (from 3 to 54 degrees). An example is illustrated in Figure 7.

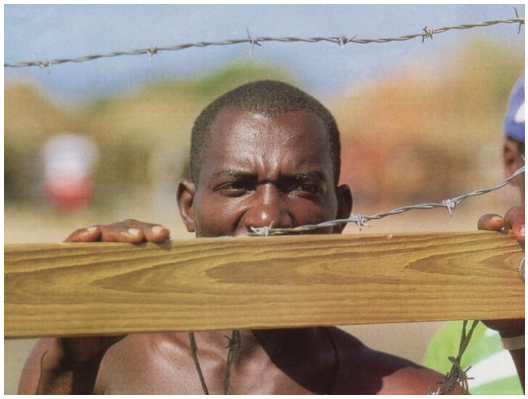

Figure 7: Stimulus image from *GAPED*.

Differently from the identification task in *b5-ref*, the goal in this case was to investigate which pictorial elements each speaker would select to describe each image, the order and structuring of these descriptions, and the lexical and syntactic choices made. To this end, the data collection was carried out in two versions - detailed and summarised text - as a means to obtain a greater degree of control over the elicited text, and to observe both discourse-level and sentence-level linguistic phenomena.

In the first task, participants were requested to fill in a text box to describe everything they could see in the scene as if they were aiding a (hypothetical) visually-impaired friend. After that, they were requested to summarise the scene contents as a single sentence (or caption) for the same purpose. For instance, a possible text description of the scene in Figure 7 may include the following example:

> '*There is a black man leaning against a barbed wire fence. He is shirtless, and he seems tired, or perhaps even sad. There seems to be a second person on his left, but he is mostly out of the picture. He is black as well, I guess, and he is wearing a blue cap.*'

A single-sentence caption for the same picture may be represented as the following example:

> '*Man looking through a fence.*'

Both *b5-text* and *b5-caption* are available in three formats: *original*, consisting of 10 files containing the set of all descriptions of each of the 10 stimulus scenes and their corresponding subject's identifiers; *per-speaker*, consisting of

151 files containing the text produced by each individual subject, and *parsed*, representing the same 10 original files with syntactic information. Generally speaking, the *original* and *parsed* formats are potentially more useful to NLG studies since they indicate what each subject wrote in response to each of the visual stimuli, whereas *per-speaker* is more useful to language understanding studies (for example, for author profiling or document classification).
As in the case of the *b5-ref* subcorpus, personality traits have also been shown to affect both the contents and the surface form of the elicited text and captions. Results of this analysis will be described elsewhere.

5. Using the b5 data

As a means to illustrate the use of the *b5* corpus data, and also to present initial reference results for future studies of this kind, this section describes a simple experiment in personality recognition from *b5-post* data. The experiment is solely focused on the recognition of the five basic personality factors, that is, individual personality facets are presently disregarded. For a more comprehensive discussion on the use of different data sources (e.g., posts, text, caption etc.) in the personality recognition task, we refer to (dos Santos et al., 2017b).

5.1. Models

Personality recognition is presently modelled as a series of five independent binary classification tasks (e.g., extrovert vs. introvert etc.) associated with each of the *Big Five* dimensions. For each of these classes, a positive label was assigned to individuals whose personality score was above the average of the group as seen in *b5-subject*, and a negative label was assigned to those whose personality score was equal or below this average.
Table 2 summarises the number of positive and negative instances for each of the five personality traits. As all classes are approximately balanced, no further re-sampling was performed.

Trait	positive	negative
Extraversion	505	514
Agreeableness	537	482
Conscientiousness	507	512
Neuroticism	548	471
Openness	533	486

Table 2: Learning instances

As learning features, we computed 64 LIWC categories (Pennebaker et al., 2001), four additional, MRC-like (Coltheart, 1981) psycholinguistic properties and further 60 dictionary attributes.
LIWC features were obtained from Brazilian Portuguese LIWC (Filho et al., 2013) by counting word categories (e.g., religion, family, money etc.). Each feature represents the number of words found in the corresponding category normalised by the length of each Facebook time line in number of words.
The four additional psycholinguistic features were obtained from (dos Santos et al., 2017a) by computing average concreteness, imageability, subjective frequency and age of acquisition scores. Each feature represents the average score of all words in the corresponding category found in each Facebook time line.
Dictionary features were obtained from Unitex-PB (Muniz, 2004) by computing word classes and a range of morphological features. Once again, each feature represents the number of words found in the corresponding category normalised by document length.
In all models, we make use of SVM classifiers with linear kernel and $\gamma = 0.1$ and $C = 1$ with 10-fold cross validation over the entire dataset.

5.2. Results

Mean precision (P), recall (R) and F1 measure (F1) scores for the five personality classification tasks are summarised in Table 3.

	positive class			negative class		
Class	P	R	F1	P	R	F1
Extraversion	0.59	0.59	0.59	0.60	0.60	0.60
Agreeableness	0.54	0.93	0.68	0.56	0.10	0.17
Conscientiousness	0.56	0.54	0.55	0.56	0.58	0.57
Neuroticism	0.55	0.95	0.69	0.59	0.09	0.15
Openness	0.57	0.67	0.61	0.55	0.44	0.49

Table 3: Personality recognition in *b5-post*

5.3. Discussion

As in existing studies of personality recognition for the English language (Mairesse et al., 2007), we notice that the Extraversion class presents the best overall results. This may suggest that this particular dimension of human personality is more evident in (Facebook) text than others.
A machine learning approach as in this example may of course be applied to many other forms of author profiling based on the *b5* corpus. These include, for instance, the classification of gender, age group and others. However, since these tasks require the additional definition of how they would be modelled in the form of a classification problem (e.g., binary, multi-class, etc.), this kind of investigation would be outside the scope of the current discussion. A number of author profiling tasks of this kind, based on *b5* data, are discussed in (Hsieh et al., 2018).

6. Final remarks

This article has described the construction of the *b5* corpus, a collection of texts produced in different communicative tasks, and accompanied by the inventories of personality of their respective authors. The *b5* corpus represents, to the best of our knowledge, the largest resource of the kind available for the Brazilian Portuguese language, and it is potentially useful for studies of computational recognition of personality traits from texts, author profiling, natural language generation based on personality traits and others. Some of these alternatives are summarised as follows.
The *b5-subject* knowledge base contains Big Five personality information, and additional attributes regarding subject's gender, age, background and others. As a result,

it may provide knowledge not only for the computational study of human personality proper as in (Mairesse et al., 2007; Farnadi et al., 2013; Nowson and Gill, 2014), but also for the study of other forms of author profiling, as in (Schwartz et al., 2013; Marquardt et al., 2014; Álvarez-Carmona et al., 2015; González-Gallardo et al., 2015; Șulea and Dichiu, 2015; Najib et al., 2015).

The *b5-post* subcorpus - containing Facebook status updates - is a textual base developed primarily for the purpose of personality recognition and author profiling in Portuguese. Studies of this kind would typically take the form of a supervised (Mairesse et al., 2007), or semi-supervised (Celli, 2012) learning task. The experiment described in the previous section is an (admittedly simple) example of the former.

The *b5-ref* subcorpus intends to support studies on machine-learning referring expression generation (REG) that take personality information into account. Studies of this kind may be seen as a possible generalisation of models of human variation for this task (Viethen and Dale, 2010; Ferreira and Paraboni, 2017). Moreover, we notice that the corpus may be also useful for studies of personality-based surface realisation and lexical choice of definite descriptions.

The *b5-text* subcorpus is potentially useful as a means to establish mapping between linguistic features and personality traits, which may guide the design of text generation models based on personality with a particular focus on multi-sentential phenomena. Given the relatively controlled domain - based on the same set of images described by all participants - *b5-text* texts make evident the different linguistic choices made by each individual. These choices, which may or may not be due to differences in personality, are observable both in surface forms and contents.

Finally, the *b5-caption* subcorpus complements the previous *b5-text* data by providing a shortened version of the same image descriptions. This corpus may be particularly useful for the study of more superficial linguistic features - such as syntactic structures and lexical choice - and their relation to personality. In addition to that, since captions may be seen as a short, single sentence summary of the larger *b5-text* text, *b5-caption* may be useful also for the development of text summarisation approaches that take the personality of the human summariser into account.

As future work, we intend to provide the semantic annotation of the stimulus scenes in *b5-text* and *b5-caption*, so that these datasets may be explored more fully in subsequent NLG studies. This work is currently in progress.

The complete *b5* corpus , currently in its version 1.7., is available under a Creative Commons Attribution 4.0 International License. The corpus may be freely downloaded[3] and reused for research purposes.

7. Acknowledgements

This work has been supported by grant # 2016/14223-0, São Paulo Research Foundation (FAPESP).

[3] https://drive.google.com/open?id=0B-KyU7T8S8bLTHpaMnh2U2NWZzQ

8. Bibliographical References

Álvarez-Carmona, M., López-Monroy, A., y Gómez, M. M., Villaseñor-Pineda, L., and Escalante, H. (2015). INAOE's participation at PAN'15: Author Profiling task. In *CLEF 2015*.

Celli, F. (2012). *Adaptive Personality Recognition from Text*. Ph.D. thesis, University of Trento.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33(4):497–505.

Dale, R. and Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In *ENLG'09*, pages 58–65, Athens.

Dan-Glauser, E. S. and Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2):468–477.

de Andrade, J. M. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Ph.D. thesis, University of Brasília.

dos Santos, L. B., Duran, M. S., Hartmann, N. S., Junior, A. C., Paetzold, G. H., and Aluísio, S. M. (2017a). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In *TSD-2017*.

dos Santos, V. G., Paraboni, I., and Silva, B. B. C. (2017b). Big five personality recognition from multiple text genres. In *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, pages 29–37, Prague, Czech Republic. Springer-Verlag.

Farnadi, G., Zoghbi, S., Moens, M.-F., and de Cock, M. (2013). Recognising personality traits using Facebook status updates. In *Proceedings of WCPR13 in conjunction with ICWSM-13*, Boston, USA. The AAAI Press.

Ferreira, T. C. and Paraboni, I. (2017). Generating natural language descriptions using speaker-dependent information. *Natural Language Engineering*, 23(6):813–834.

Filho, P. P. B., Aluísio, S. M., and Pardo, T. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *proc. of STIL-2013*, pages 215–219, Fortaleza, Brazil.

Gatt, A., van der Sluis, I., and van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proc. of ENLG-07*.

Gauthier, I. and Tarr, M. J. (1997). Becoming a greeble expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682.

Goldberg, L. R. (1990). An alternative description of personality: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.

González-Gallardo, C., Montes, A., Sierra, G., Núñez-Juárez, J., Salinas-López, A., and Ek, J. (2015). Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams. In *CLEF 2015*.

Hsieh, F. C., Dias, R. F. S., and Paraboni, I. (2018). Author profiling from facebook corpora. In *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*, Miyasaki, Japan. ELRA.

Iacobelli, F., Gill, A. J., Nowson, S., and Oberlander, J.

(2011). Large scale personality classification of bloggers. In *ACII (2)*, pages 568–577. Springer.

John, O. P., Donahue, E., and Kentle, R. (1991). The Big Five inventory - versions 4a and 54. Technical report, Inst. Personality Social Research, Univ. California.

Krahmer, E. and van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Lan, A. G. J. and Paraboni, I. (2018). Definite description lexical choice: taking speaker's personality into account. In *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*, Miyasaki, Japan. ELRA.

Mairesse, F. and Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.-F., Davalos, S., Teredesai, A., and de Cock, M. (2014). Age and gender identification in social media. In *CLEF 2014 Working notes*, pages 1129–1136, Sheffield, UK.

Muniz, M. C. M. (2004). A construção de recursos linguístico-computacionais para o Português do brasil: o projeto de Unitex-PB. Master's thesis, USP São Carlos.

Najib, F., Cheema, W. A., and AdeelNawab, R. M. (2015). Author's traits prediction on Twitter data using content based approach. In *CLEF 2015 Working Notes*.

Nowson, S. and Gill, A. J. (2014). Look! who's talking?: Projection of extraversion across different social contexts. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 23–26, Orlando, FL, USA. ACM.

Paraboni, I., Galindo, M., and Iacovelli, D. (2017a). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 51(2):439–462.

Paraboni, I., Lan, A. G. J., de Sant'Ana, M. M., and Coutinho, F. L. (2017b). Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, 43(2):451–459.

Paraboni, I., Monteiro, D. S., and Lan, A. G. J. (2017c). Personality-dependent referring expression generation. In *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, pages 20–28, Prague, Czech Republic. Springer-Verlag.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

Righi, G., Peissig, J. J., and Tarr, M. J. (2012). Recognizing disguised faces. *Visual Cognition*, 20(2):143–169.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., and Ungar, L. H. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*, 8(9):e73791.

Soto, C. J. and John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, 43(1):84–90.

Șulea, O.-M. and Dichiu, D. (2015). Automatic Profiling of Twitter Users Based on Their Tweets. In *CLEF 2015*.

Viethen, J. and Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In *Proc. of the Australasian Language Technology Association Workshop 2010*, pages 81–89, Melbourne.