Journal of the
Brazilian Computer Society

RESEARCH Open Access

Personality-dependent content selection in natural language generation systems

Ricelli M. S. Ramos, Danielle S. Monteiro and Ivandré Paraboni*

*Correspondence: ivandre@usp.br
University of São Paulo, School of Arts, Sciences and Humanities, Av Arlindo Bettio 1000, São Paulo, Brazil

Abstract

This paper focuses on the computer side of human-computer interaction through natural language, which is the domain of natural language generation (NLG) studies. From a given (usually non-linguistic) input, NLG systems will in principle generate the same fixed text as an output and in order to attain more natural or human-like interaction will often resort to a wide range of strategies for stylistic variation. Among these, the use of computational models of human personality has emerged as a popular alternative in the field and will be the focus of the present work as well. More specifically, the present study describes two machine learning experiments to establish possible relations between personality and content selection (as opposed to the more well-documented relation between personality and surface realisation), and it is, to the best of our knowledge, the first of its kind to address this issue at both macro and micro planning levels, which may arguably pave the way for the future development of more robust personality-dependent systems of this kind.

Keywords: Natural language generation, Content selection, Personality traits, Big five

Introduction

Natural language generation (NLG) systems produce text from (usually) non-linguistic input and are central to the development of realistic, psychologically plausible human-computer communication that does not resort to pre-defined or 'canned' text. Applications include the generation of textual summaries from neonatal intensive care data [1], patient history and nurse reports [2–4], personalised smoking cessation letters [5], weather forecasts [6], dialogue and narrative text [7, 8], poetry [9], image captions [10, 11], and many others.

The design of a NLG system often follows a general 3-stage pipeline architecture as in [12], comprising document and sentence planning (also known as macro and micro planning), and surface realisation. These components are illustrated in Fig. 1.

Starting from a high-level communicative goal of describing a given input meaning as text, a typical NLG system will first build up a document plan that provides the set of

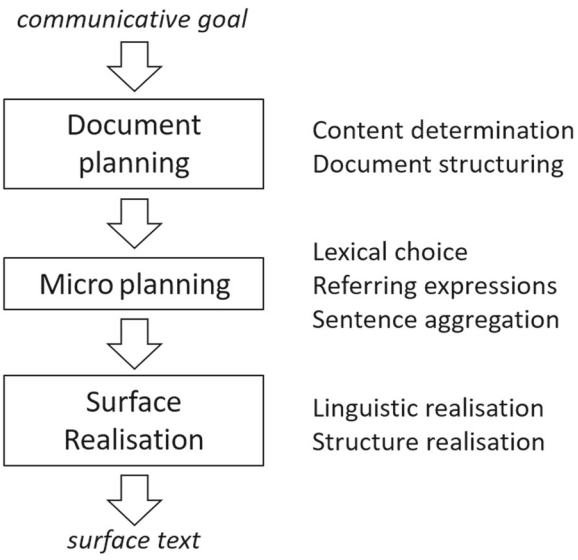

Fig. 1 The NLG pipeline in [12]

contents (e.g. discourse objects and their semantic properties) to be presented in the output and some form of high-level structure (e.g. content ordering, rhetorical relations) Next, the document plan is refined into a series of abstract sentence representations in which specific words are selected to express plan concepts (a task known as lexical choice [12]), context-dependent referring expressions are fully specified, and concepts are combined into abstract sentence units. Finally, sentence units are rendered in a target natural language and structured according to an appropriate grammar formalism. For a detailed discussion on this architecture, we refer to [12].

Customised natural language generation

NLG systems may in principle produce always the same fixed output text for a given input meaning. However, systems that aim to generate text in a more natural or human-like fashion will often implement a wide range of strategies to model some form of stylistic variation. Among these, the use of computational models of *human personality* has emerged as a popular alternative in the field.

Of particular interest for the present work, we will consider the use of the Big Five model of human personality [13] in NLG. The model is based on the assumption that differences in personality are revealed by the way individuals express themselves in natural language and, given its linguistic motivation, is not surprising that the Big Five model has been applied to a wide range of studies in both natural language understanding [14–16] and generation [17, 18] alike.

The Big Five model comprises five fundamental dimensions of personality: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, and *openness to experience*. To appreciate the role of these personality traits in language production, let us consider the task of producing a simple text description of an input scene as in Fig. 2.

In a situation of this kind, different human speakers may produce a large number of alternative text descriptions. For instance, the following are examples of how the example scene may be described by two speakers with different degrees of agreeableness.

Fig. 2 A scene from GAPED [41]

speaker1 (higher agreeableness):
"*It is the photo of a girl of around 12 and her brother, probably 6 years old. They are on a dirt road. In the background it is possible to see vegetation, and in the front of them there are some scattered wires for building fences. The girl wears a long brownish dress with long sleeves, and red slippers. The boy wears a jumper and trainers.*"

speaker2 (lower agreeableness):
"*They are two children, a boy and a girl, who seem to be poor by the clothes they are wearing, and who are in a place that seems to be rural because of the dirt road and the bushes, and are close to rubbish and building materials.*"

The two texts in this example are obviously different in a number of ways. At the surface level, for instance, we notice that the speakers use different wordings to describe the same pictorial element. This is the case, for instance, of the background vegetation in the scene, described as 'vegetation' by speaker1 but described as 'bushes' by speaker2. This difference—which is largely an issue of lexical choice—has been a primary focus in personality-dependent NLG research [19], and it is consistent with the lexical motivation at the very core of the Big Five model [13].

Lexical choice is however only one among many differences between the two example descriptions. In particular, we notice that the two speakers chose to select different *contents* to appear in these text descriptions. Thus, for instance, speaker1 seems somewhat more focused on the two main human characters in the scene and uses more colour information than speaker2. From a NLG perspective, differences in meaning as in these examples are not surface realisation issues, but rather semantic *content selection* (CS) decisions to be dealt with at the early stages of the pipeline architecture.

The distinction between wordings and meanings is of course debatable. For instance, are 'small' and 'short' simply different wordings to describe the same meaning (i.e. the boy's height in the picture), or do they actually convey different meanings? Leaving these difficulties aside, however, the distinction between text semantics and surface form is crucial for practical NLG, and it is central to the present study on personality-dependent CS.

Objectives

The main objective of the present work is to show that, in personality-dependent NLG systems, personality traits of the target speaker may influence not only surface realisation, but also content selection. In other words, we would like to show that personality traits affect not only 'how we say it' (at the surface level) but also what we actually choose to say in the first place (at a deeper semantic level.)

Moreover, we also would like to show that the use of personality information enables a CS model to make more accurate predictions at both macro and micro planning levels. To this end, we shall consider two instances of the CS task: the more coarse-grained kind of CS performed at the document planning stage of the NLG pipeline, hereby called an instance of *discourse-level* CS, and the more fine-grained CS task performed at the microplanning stage to produce referring expressions in a particular point in the discourse (e.g. 'the girl', 'she', 'the tallest child'), hereby called *reference-level* CS (known in the NLG field as the referring expression generation (REG) task.)

In Fig. 2, an example of discourse-level CS task would consist of deciding which characters or objects should be mentioned in a text description. In the NLG architecture, discourse-level CS is generally driven by the communicative goals provided as an input to the system (e.g. the goal of describing a picture according to a given personality profile.)

An example of reference-level CS, on the other hand, would consist of providing an unambiguous referring expression to enable the identification of a particular target (e.g. 'the girl on the left'.) This task is driven by the need to produce uniquely identifying referring expressions - often in the form of definite descriptions as in this example - of the intended target in a particular context (e.g. by taking into account both the visual scene and/or the entities mentioned in the recent discourse.)

Although discourse- and reference-level CS arguably address a similar underlying issue, in what follows the two tasks are discussed separately in two independent experiments. In both cases, however, the experiments make use of controlled text produced by a single group of speakers, as provided by a corpus of text descriptions labelled with personality information about their authors.

The rest of this paper is structured as follows. The "Related work" section briefly addresses existing work in NLG content selection from both macro and micro planning perspectives. The "Experiment 1: Personality-dependent discourse-level CS" section presents the experiment in personality-dependent discourse-level CS, from data collection to model design and evaluation. The "Experiment 2: Personality-dependent reference-level CS" section follows a similar structure to address the second experiment, devoted to personality-dependent reference-level CS. The "Final remarks" section draws a number of conclusions and hints at future work.

Related work

Examples of personality-dependent NLG systems are few, and even when a system does address the issue of how personality information may be embedded in language generation the focus is usually on the surface realisation or lexical choice tasks rather than content selection. Among the existing studies of this kind, the PERSONAGE system [20] and its extensions are, to the best of our knowledge, the most complete examples of text-generating systems that take personality information into account. PERSONAGE and a few other examples of personality-based NLG are briefly reviewed in the

"Content selection for document planning" section from a macro planning perspective. This is followed by a more detailed discussion regarding the micro planning task of Referring Expression Generation (REG) in "g" section.

Content selection for document planning

Given a mass of usually non-linguistic input data, content selection as performed in the macro planning stage of the NLG architecture - hereby called discourse-level CS - consists of selecting the meanings to be represented as text in the subsequent stages of the system pipeline [21]. The issue of personality-dependent discourse-level CS is however little discussed in the NLG literature, perhaps based on the observation that models such as the Big Five [13] are largely focused on the relation between personality and word choice.

The work in [20] introduces PERSONAGE, a first attempt to develop a fully-functional personality-dependent NLG system in the restaurant recommendation domain. PERSONAGE supports a range of stylistic variations that may be controlled by personality information provided as an input. The work focuses on the effects of the Big Five *Extraversion* trait over the output text, and investigates how differences in personality are perceivable by human readers.

PERSONAGE makes use of machine learning methods to map personality traits to generation decisions that affect the output text. Most of the resulting variation is related to sentence structuring and word choice. For instance, the system favours the generation of longer sentences with a higher number of negations, and uses more tentative words when an introvert profile is selected.

Since the input to the system consists of user-determined communicative goals, content selection is mainly focused on structuring the given input in order to maximise the perception of differences in system personality. Generation decisions that may vary along the *Extraversion* dimension include degree of verbosity, number of repetitions and concessions, among others.

Many subsequent studies were developed as extensions of PERSONAGE. These include a wide range of improvements on the system architecture and support to additional personality traits [17], and its application to other domains such as gossip generation [22], computer game dialogues [23], creative writing [24], storytelling [25], gesture generation [26] and customer feedback generation [18], among others. Generally speaking, however, the relation between personality and content determination is not the focus of any of these studies.

Content selection for referring expression generation

A second and much more fine-grained form of content selection is the case of content selection for referring expression generation (REG), which is performed at the micro planning stage of the NLG pipeline. REG is concerned with the generation of uniquely identifying definite descriptions (more generally known as referring expressions) of a given target object, so that the generated descriptions resemble those that would have been produced by human speakers. For instance, let us consider the goal of describing the target *o*4 in Fig. 3[1].

[1] Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon Univ. Funding provided by NSF award 0339122.

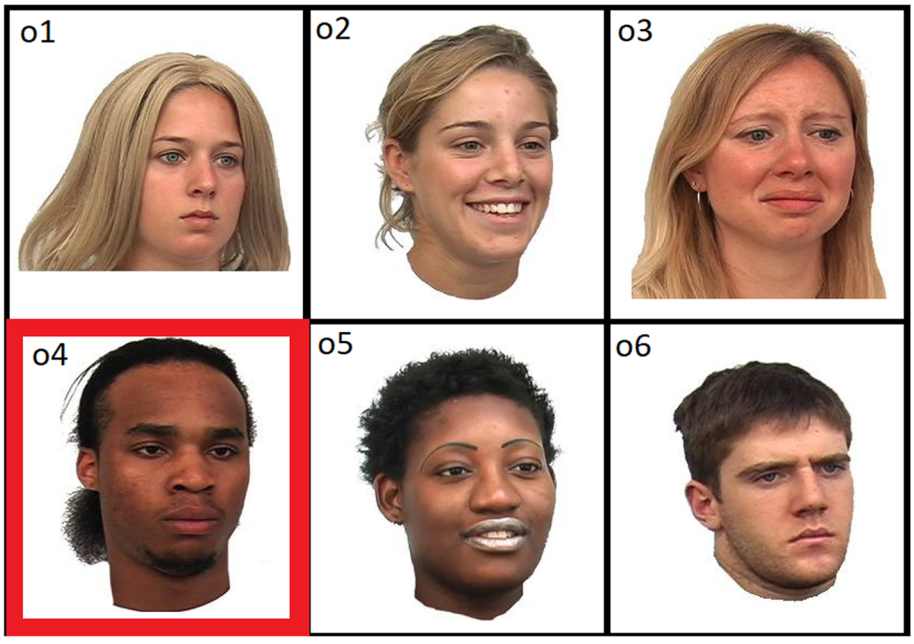

Fig. 3 An example of referential context built from Face Place images [49]. Object labels o1..o6 were added for ease of discussion

Given a target object that we intend to describe, and a context set containing a number of distractor objects (i.e. the other characters in the example scene), the goal of a REG algorithm is to select the contents to compose an unambiguous description of the target object, either in atomic (e.g. 'the dark man', 'the guy with a pony tail') or relational [27] form (e.g. 'the man below a girl on the left side'.)

Referring expressions are ubiquitous in text, and although in this example we use a visual context for ease of discussion, REG algorithms are actually required to generated every referring expression that occurs in the text, including, for instance, references to any discourse object with respect to the set of objects recently mentioned in an ongoing conversation (e.g. 'the problems what we discussed yesterday', 'the second and third ones', 'these issues'). Reference-level CS is therefore a mandatory task in any sufficiently complex NLG system and, accordingly, a well-studied research topic in the field [28, 29].

The choices of which facts are selected to make a uniquely identifying description are largely determined by its referential context, and REG is largely driven by the need to prevent ambiguity within such context while avoiding the generation of overly long or otherwise redundant descriptions. More formally, reference-level CS task takes as an input a target r to be distinguished from a set of distractor objects within a given context C. Objects are usually modelled as sets of properties represented as (*attribute*-value) pairs, as in (*ponytail*-yes.) The goal of a REG algorithm is to produce a set L of properties that are true of r such that L distinguishes r from every distractor in C [28]. The output description L may be subsequently realised as a definite or indefinite description. For instance, an output description, $L = \{$*gender*-male, *ponytail*-yes$\}$ could be realised as 'the guy with a pony tail'.

One of the best-known approaches to REG is the Incremental algorithm in [30]. In this approach, attributes are considered for selection according to a domain-dependent

list of preferences P and provided that they are discriminatory, that is, provided that the selected attributes rule out at least one distractor object in the context. When an attribute a is selected, the corresponding distractor objects are removed from C. For instance, selecting a property (*gender*-male) rules out all distractor objects whose *gender* is female. The algorithm terminates when C becomes empty or when all attributes in P have been attempted.

The incremental approach and many of its successors are generally concerned with the generation of a single, fixed description for the given input. By contrast, more recent approaches as in [31, 32] have addressed the issue of *human variation* in REG as well. These methods, however, rely on a set of pre-recorded examples of referring expression produced by every speaker under consideration and take as an input a unique identifier of the target speaker to generate highly specialised descriptions. As a result, speaker-dependent REG may be of limited practical use unless suitable (linguistic) training data is available.

One possible way of adding human variation to the output descriptions generated by a REG algorithm without resorting to a large among of linguistic examples as training data is by assuming that *personality* may play a role in the content selection of referring expressions as well. Based on this observation, the study in [33] addressed the issue of how personality may affect referential overspecification, that is, the use of additional information in the referring expression beyond what is strictly required for disambiguation (e.g. the affective information (upset) in 'the guy with a pony tail, who looks upset' is redundant.) The study however falls short of providing a full personality-dependent REG algorithm, focusing instead on the question of how to modify a description previously generated by a standard REG approach to accommodate a certain level of personality-dependent variation, and we are not aware of other personality-dependent studies that have addressed the issue of reference-level CS in more detail.

Experiment 1: Personality-dependent discourse-level CS

This section addresses the issue of how a given input scene may be described by speakers with different personality types. From a semantic representation of a scene as in Fig. 2 (i.e. objects and their properties), discourse-level CS should in principle contemplate two questions: deciding *which objects* should be mentioned (e.g. whether to mention the girl, the boy, or possibly both) and deciding *which facts* (or semantic properties) about these objects should be selected (e.g. whether to mention the fact that the girl seems sad, or the fact that her shoes are red). For reasons to be discussed in the next section, however, in what follows, we pay no regard to the former, and we will focus instead on the computational task of selecting object properties.

The current approach to discourse-level CS makes use of supervised machine learning methods to select scene properties based on a target personality. After data preparation, a series of classifiers are built from training data to predict whether each individual property should be selected. Next, individual predictions made by every classifier are combined to produce a set of properties that represents the semantic contents of a possible text description of the input scene. These steps are discussed individually in the next sections.

Data

The present work makes use of the *b5* corpus [34] of Brazilian Portuguese texts produced by participants in a controlled data collection task. The texts were elicited in a number of communicative tasks and subsequently labelled with Big Five scores obtained from a BFI-44 personality inventories [35] filled in by the participants themselves. The corpus has been previously taken as the basis to a number of studies in personality recognition [36–39] and author profiling [40] from text. In what follows, we will focus on the *b5-text* (sub)corpus of scene descriptions.

The *b5-text* dataset was primarily collected for the study of issues of personality-dependent content selection and surface realisation. The data consist of scene descriptions elicited from visual stimuli taken from *GAPED* [41], a collection of images classified by valence and normative significance, which are designed so as to arouse different degrees of emotional response. An example of one such image was presented in the previous Fig. 2.

Data collection made use of ten stimulus images with degrees of valence in the 3–54 range as provided by GAPED. For each image, participants of an in-person data collection task were requested to describe everything that they could see in the scene, as if helping a (hypothetical) visually-impaired friend. The text descriptions labelled as speaker1 and speaker2 in the "Introduction" section are translated examples of actual instances of the *b5-text* corpus. Notice however that, since the underlying meanings change across scenes, each image is to be treated independently in our experiments, that is, the text descriptions in the corpus effectively make ten unrelated datasets. Thus, for instance, the observation that a particular scene shows a smiling person is in principle unrelated to the observation that there is a smiling person in another, or even in the same scene. The two smiles simply represent two different facts, and the use of ten unrelated input scenes is simply an attempt to increase the likelihood of finding effects of personality on content selection, which may be easily influenced (or even determined) by perception. In particular, we notice that certain objects may be more salient—and possibly more prone to mention—than others and that visual salience alone may obscure any effects that personality might have on content selection. Thus, given that perception skill levels may vary across individuals, trying to establish beforehand which single scene might be more indicative of differences across multiple personality types would have been unhelpful.

The *b5-text* corpus comprises 1510 scene descriptions produced by 151 participants. However, given the nature of the underlying language production task, in which participants were instructed to describe *all* objects in each scene, differences in personality are not truly reflected in their choices of scene objects, and the few existing differences across speakers are insufficient for our current purposes. For that reason, our work will disregard the issue of which scene objects (which are usually realised as nouns in the text) are to be selected and will focus instead on the choices of facts (or semantic properties) about these objects (which are usually realised as adjectives.) Thus, the present experiment should be more appropriately described as a first step towards more comprehensive (personality-dependent) document planning studies rather than a fully functional CS module.

An investigation of content selection choices gives rise to the question of how to define a content unit, that is, which semantic properties should be modelled based on the existing text. As discussed above, we shall focus on properties that are realised in surface form

Table 1 Properties with 20 or more occurrences in each set of scene descriptions

Scene	#	Adjective-noun pairs
01	6	{red-sandals (46), brown-dress (37), long-dress (26), blue-pants (22), tall-girl (22), older-girl (22)}
02	4	{black-canvas (74), white-bucket (65), large-bucket (30), small-child (28)}
03	5	{fat-lips (89), white-bedspread (59), red-details (56), red-lips (29), black-baby (22)}
04	2	{black-person (75), raised-arms (35)}
05	8	{sitting-people (38), green-coat (38), white-hair (28), white-trainers (25), sitting-lady (24), blue-coat (21), red-coat 20, black-shirt (20) }
06	8	{pink-bedspread (79), blue-blanket (74), blue-coat (67), dirty-coat (55), black-baby (54), dirty-clothes (32), sick-child 29, dark-skin (19)}
07	0	{-}
08	5	{black-man (111), blue-cap(35), barbed-wire (25), green-shirt (24), short-hair (22)}
09	9	{yellow-shorts (51), plaid-shorts (45), blue-shirt (37), barefoot-man (36), elderly-man (34), crossed-legs (29), grey-shorts (28), red-tiles (28), purple-hat (21)}
10	4	{dry-grassland (40), elderly-woman (30), red-pants (30), red-hat (22)}

as adjectives (e.g. 'red', 'sad'). More specifically, in order to keep the necessary distinction in meanings across scenes and across subjects within the same scene (e.g. the property of 'being dark' has different meanings in 'dark hair' and 'dark skin'), each adjective in the text corpus is labelled with a unique identifier representing its underlying concept and the associated subject (a head noun or its pronoun substitute), as in, e.g. 'dark_hair'. Tuples of this kind make the input semantic properties to our model.

In order to obtain the list of properties represented in a given text description, a combination of automatic and manual[2] annotation tasks was performed. First, adjective-noun and adjective-pronoun pairs were extracted from the text descriptions of every scene with the aid of the PALAVRAS syntactic parser of Portuguese [42]. This was followed by manual revision to correct common parsing errors (e.g. adjectives attached to the wrong head) and also to replace pronouns for their actual noun antecedents (e.g. 'they' may be replaced by 'shoes'). Next, nouns denoting the same concept (e.g. 'shoes' and 'trainers') were clustered together when applicable. Clustering was generally straightforward as it is usually clear from the scenes which pictorial elements are referred to in each text description. Finally, all adjective-noun pairs with fewer than 20 occurrences in the dataset were discarded.

From the ten stimulus images, a set of 51 properties above the minimum (20) threshold were identified. These properties are listed in Table 1, accompanied by their actual number of instances (between brackets). In the case of scene 7, we notice that all existing properties fell below the minimum, and for that reason, this scene will be disregarded.

In this dataset, in which several classes barely reach the 20-instance minimum, we notice that data sparsity is a major concern. Moreover, as discussed in the previous section, we are aware that we may not necessarily find (major) personality effects on content selection for every scene. These issues will be further discussed in the "Evaluation" section.

[2] Performed by two annotators and subsequently revised by a third judge.

Finally, text descriptions were divided into training and test sets in a balanced 80:20 split so that descriptions produced by every speaker appeared in both subsets. As a result, 1200 training descriptions and 300 test descriptions were created. Personality scores associated with each description were obtained from the *b5-text* corpus as discussed in [34].

Models

We envisaged a discourse-level CS model—hereby called *PersonalityDoc*—based on a series of classifiers that takes as an input a scene conveying a set of objects D and their properties, a target object r for which a property p is known to be true, and a target Big Five profile b. The output is a prediction of whether p should be selected as part of a text description of D as uttered by a speaker of personality b.

Individual classifiers were built for each of the 51 domain properties under consideration as defined in the previous section. Thus, for instance, one classifier is intended to predict whether to select the *smiling-girl* property in a given context, another predicts whether to select *red-shoes* and so forth. In all cases, the set of learning features consists of the scene and semantic property identifiers (which in turn correspond to adjective-noun pairs), and the five scalar values representing the personality scores of the target speaker.

Learning instances were computed as follows. For every text description of every scene, a positive instance of class p is created whenever p occurs in a text, and a negative instance of class p is created otherwise. We notice that the resulting dataset is prone to class imbalance, with an average 2:1 negative-positive instance ratio.

Evaluation

Assessing *PersonalityDoc* is complicated by the fact that no similar systems are immediately available for use as a baseline. However, given that our dataset is heavily imbalanced, we will use this to our advantage and evaluate *PersonalityDoc* against a majority class baseline.

Both models were built using linear support vector machine (SVM) with optimal parameter values obtained by performing grid search on the training dataset. The choice for SVM was motivated by the relatively small size of the current dataset and by positive results obtained in related tasks [43].

As a means to address the issue of class imbalance, *PersonalityDoc* makes use of the Synthetic Minority Over-sampling Technique SMOTE [44] with k=5 nearest neighbours. The majority class model, on the other hand, does not make use of oversampling since this would result in a much weaker baseline. The comparison between these two models gives rise to the following research hypothesis:

> *h1*: The use of personality information about a target speaker enables a document-level CS model to select document contents that resemble more closely the choices made by humans if compared to a similar model that does not have access to personality information.

This hypothesis will be verified by comparing the contents selected by *PersonalityDoc* with the contents selected by the majority class baseline from the same input. We expect that, on average, the contents selected by *PersonalityDoc* will resemble human choices more closely than those selected by the baseline as measured by mean Dice coefficients [45].

Table 2 Mean precision (P), recall (R), and F1-measure (F) 10-fold cross validation results on training data, per scene

		Majority class baseline			Personality-dependent CS		
Scene	Classes	P	R	F	P	R	F
01	6	0.81	1.00	0.89	0.83	0.99	**0.90**
02	4	0.67	1.00	0.80	0.82	0.99	**0.90**
03	5	0.69	1.00	0.81	0.79	0.98	**0.88**
04	2	0.57	1.00	0.72	0.79	0.99	**0.87**
05	8	0.82	1.00	0.90	0.82	1.00	0.90
06	8	0.67	1.00	0.79	0.79	0.98	**0.88**
08	5	0.80	1.00	0.89	0.81	0.99	0.89
09	9	0.76	1.00	0.86	0.78	0.99	**0.87**
10	4	0.80	1.00	0.89	0.81	0.99	0.89
Mean	5.1	0.73	1.00	0.84	0.80	0.99	**0.89**

Best F1 results are highlighted

Classification results

Before discussing the actual CS results in the next section, we start by assessing the predictions made by the individual classifiers using 10-fold cross validation over the entire dataset. Results obtained by the baseline and personality-dependent models (except for scene 07) are shown in Table 2.

From these results, we notice that, on average, personality-based CS outperforms the baseline method in most contexts under consideration, or it is at least equally effective. A more detailed view of the same data is presented in Table 3, in which results for the most frequent properties of each of the nine sets of scene descriptions are presented next to the number of positive and negative instances of each class.

Once again, we notice that personality-dependent classifiers generally outperforms the majority class baseline for most input scenes and properties.

Content selection results

Table 4 presents mean Dice scores obtained by *PersonalityDoc* and baseline models applied to the generation of the test descriptions for each of the nine input scenes in the corpus.

From these results, we notice that overall Dice scores obtained ny *PersonalityDoc* are higher than those obtained by the baseline model. The difference is significant according to a Wilcoxon signed rank test (W =-5127, z =-5.57, p <0.001). The use of personality information about a target speaker enables the CS model to select property sets that resemble more closely those that would be selected by humans. This offers support to hypothesis $h1$.

Experiment 2: Personality-dependent reference-level CS

In this section, we further this issue of personality-dependent CS by zooming into the content selection task of individual referring expressions, as in 'the guy with a ponytail', 'the dark-skinned man who looks upset', and so on. This issue is the domain of the referring expression generation (REG) NLG subtask, addressed in the micro planning stage of the NLG pipeline [12].

Table 3 Mean precision (P), recall (R), and F1-measure (F) 10-fold cross validation results on training data for the most frequent properties per scene

			Majority class baseline			Personality-dependent CS		
Scene	Property	Inst.+/−	P	R	F	P	R	F
01	red-sandals	46/105	0.70	1.00	0.82	0.82	0.96	**0.88**
01	brown-dress	37/114	0.76	1.00	0.86	0.86	0.99	**0.92**
01	long-dress	26/125	0.83	1.00	**0.91**	0.84	0.98	0.90
02	black-canvas	74/77	0.51	1.00	0.68	0.81	0.99	**0.89**
02	white-bucket	65/86	0.57	1.00	0.73	0.76	0.98	**0.86**
02	large-bucket	30/121	0.80	1.00	0.89	0.86	1.00	**0.92**
03	fat-lips	89/62	0.56	1.00	0.71	0.77	0.99	**0.87**
03	white-bedspread	59/92	0.62	1.00	0.76	0.74	0.99	**0.85**
03	red-details	56/95	0.54	1.00	0.78	0.75	0.99	**0.85**
04	black-person	75/76	0.52	1.00	0.69	0.84	0.97	**0.90**
04	raised-arms	35/116	0.62	1.00	0.76	0.73	1.00	**0.84**
05	sitting-people	38/113	0.75	1.00	0.86	0.86	0.96	**0.81**
05	green-coat	38/113	0.75	1.00	0.86	0.77	1.00	**0.87**
05	white-hair	28/123	0.81	1.00	**0.90**	0.77	1.00	0.87
06	pink-bedspread	79/72	0.52	1.00	0.68	0.79	0.99	**0.88**
06	blue-blanket	74/77	0.51	1.00	0.68	0.83	0.97	**0.89**
06	blue-coat	67/84	0.56	1.00	0.71	0.77	0.99	**0.87**
08	black-man	111/40	0.72	1.00	0.84	0.86	0.96	**0.91**
08	blue-cap	35/116	0.77	1.00	**0.87**	0.72	1.00	0.84
08	barbed-wire	25/126	0.83	1.00	0.91	0.85	1.00	**0.92**
09	yellow-shorts	51/100	0.54	1.00	0.70	0.75	0.98	**0.85**
09	plaid-shorts	45/106	0.70	1.00	0.82	0.80	0.99	**0.89**
09	blue-shirt	37/114	0.75	1.00	0.86	0.77	0.97	0.86
10	dry-grassland	40/111	0.74	1.00	0.85	0.83	0.99	**0.90**
10	elderly-woman	30/121	0.80	1.00	0.89	0.83	0.99	**0.90**
10	red-pants	30/121	0.80	1.00	**0.89**	0.75	1.00	0.86
	Mean	51/100	0.68	1.00	0.80	0.80	0.99	**0.88**

Best F1 results are highlighted

Table 4 Mean Dice coefficients and standard deviation per input scene

	Baseline		PersonalityDoc	
Scene	Mean	SD	Mean	SD
01	0.00	0.00	**0.27**	0.32
02	0.00	0.00	**0.16**	0.30
03	0.42	0.35	0.42	0.35
04	0.00	0.00	**0.30**	0.41
05	0.00	0.00	**0.14**	0.24
06	0.25	0.28	**0.32**	0.24
08	**0.56**	0.41	0.34	0.29
09	0.00	0.00	**0.35**	0.28
10	0.00	0.00	**0.21**	0.32
Mean	0.14	0.12	**0.28**	0.31

Best results are highlighted

The current experiment will once again make use of supervised machine learning methods to select properties based on a target personality. Unlike the discourse-level CS experiment, however, in the present case, we go one step further and use the selected properties in an actual REG algorithm in order to generate personality-dependent object descriptions.

Data

Participants of the data collection task in [34] were also requested to produce a number of referring expressions under controlled circumstances, resulting in the *b5-ref* corpus of definite descriptions. The corpus was built by making use of a standard referential task—of the kind commonly found in data collection tasks for REG (e.g. [46–48])—in which participants were presented with a series of 12 referential contexts built from the Face Place image database [49] as in the previous Fig. 3 and were requested to uniquely describe a particular target by completing a sentence in the form 'The person highlighted in red is the...' for each scene.

Each context image displayed six human faces with various physical and affective traits. Different situations of reference required participants to deal with various levels of ambiguity and properties with different degrees of salience (e.g. a scene containing a single, prominent smiling person, or several characters with a similar hair style). For further details, we refer to [34].

A corpus of 1822 word strings representing face descriptions was obtained. In order to use the data in our current CS study, both input scenes and their descriptions were semantically annotated by two judges according to a 27-attribute annotation scheme corresponding to the most frequent information observed in the elicited data.

Unlike the scenes in the *b5-text* domain considered in the previous experiment, we notice that *b5-ref* referential contexts are considerably more homogeneous, that is, all contexts depict human faces in a similar fashion. With few exceptions, the corresponding descriptions are mostly limited to a small set of possible attributes (e.g. gender, race), which in principle allows us to model a single set of more general properties for all contexts (as opposed to modelling scene-specific properties). For instance, a single attribute *smile* is taken to represent any instance of smile (or lack of it) for any human character in all the 12 contexts, which in turn enables the use of standard general-purpose REG algorithms as in [30] and others as discussed below.

The values of certain attributes (e.g. *gender*, *race*) were obtained automatically from the meta data available from Face Place [49]. Others, by contrast, were annotated by choosing the value chosen by the majority of speakers was selected. Table 5 summarises the ten most frequent attributes in the corpus. From these statistics we notice that *gender* information is included in nearly all (94%) descriptions. This suggests that the role of personality in selecting *gender*, if any, is comparatively small. For that reason, in the experiment to follow, the *gender* attribute will always included in the generated descriptions regardless of the input personality traits provided.

We notice also that both *isYoung* and *eyebrows* have only one possible value each. Attributes of this kind—which have no discriminatory power (e.g. because all stimuli images depicted people who have eyebrows and who are reasonably young)—are often disregarded in standard REG [30, 50]. Given their ubiquity in real language use, however,

Table 5 Most frequent referential attributes found in the corpus descriptions

Attribute	Possible values	Instances
Gender	{male,female}	1707
Race	{asian,black,caucasian}	794
Smile	{yes,no}	784
isYoung	{yes}	705
hair.colour	{dark,blonde}	633
hair.length	{short,long }	434
Emotion	{positive,negative,neutral}	266
eye.colour	{light,dark}	191
Ponytail	{yes,no}	174
Eyebrows	{other}	156

these attributes will be kept as part of the present CS model, and we expect our REG algorithms to handle them appropriately.

The annotated descriptions—which are represented as sets of attribute-value pairs—were divided into training and test sets in a balanced fashion to ensure that descriptions produced by every speaker appeared in both subsets in similar proportions. To this end, 20 random descriptions produced by every speaker were kept as training data, and the remaining two descriptions were kept as test data. One thousand five hundred eighty-two training descriptions and 240 test descriptions were obtained in this way, roughly corresponding to a 83:17 split.

Models

We follow [32] and others and implement a machine learning REG model based on a series of individual classifiers. More specifically, we built a binary decision-tree classifier for every attribute under consideration (except *gender*, as discussed in the previous section.) Thus, for instance, a classifier predicts whether to select the *hair.colour* attribute in a given context, and a separate classifier predicts whether to select the *smile* attribute and so forth. However, since many of the 27 possible attributes available from the b5-ref corpus annotation (cf. the "Data" section) are infrequent, in what follows, we shall focus on the subset of the ten most frequent attributes described in the previous Table 5, which correspond to over 81% of all attributes that appear in this domain.

Each binary classifier—which decides whether a given attribute a should be selected or not to appear in a referring expression under construction—may take as an input two kinds of features, hereby called context and personality features. These are discussed in turn as follows.

As in standard REG, the present work relies heavily on context features as an input. Features of this kind are intended to represent the referential context within which the communication is taking place and, accordingly, are computed from the underlying input scene specification. In the present study, for every possible attribute of the target object, two kinds of context features are computed: discriminatory power (as defined in [30]) and average attribute frequencies.

Given an attribute a of the target object, discriminatory power represents the number of distractor objects that would be ruled out should a be selected, and it is intended to motivate the usefulness of selecting a for the purpose of disambiguation. For instance, in a scene where all objects have the same colour, selecting the *colour* attribute would not help

disambiguate the reference, that is, *colour* has zero discriminatory power in this context. By contrast, in a scene that includes n objects whose colour is different from the colour of the target object, the discriminatory power of *colour* equals n.

As a second kind of context feature, we also consider the average frequency of a in that particular scene as seen in training data. Features of this kind are intended to model (in a machine learning setting) context-dependent preferences not unlike the preferred list of attributes parameter in the incremental approach [30] and may be thought of as a means to model domain preferences that cannot be explained by the need to avoid ambiguity alone. This may be the case, for instance, of the well-known human preference for colour even in contexts in which colour has little or no discriminatory power [51].

Finally, in addition to the above context features, the present model also considers five personality features representing the Big Five traits associated with the speakers who produced each description. Each of these five features—extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience—is modelled as a scalar value obtained from the personality-labelled corpus (cf. the "Data" section).

For every training description L produced by a human speaker s in a context C, context and personality features are combined to make a number of learning instances as follows. Assuming that the model will be limited to the ten most frequent target attributes in the domain as discussed in the "Data" section, the full feature set will consist of 25 features: ten context features representing the discriminatory power of each target attribute (e.g. the fact that *gender* rules out four (female) distractor objects), ten context features represent their average frequencies in the training data (e.g. modelling the fact that *emotion* is the least frequently chosen attribute in this particular scene), and five personality features representing the Big Five traits of the target speaker.

REG algorithm

Based on the machine learning REG strategy discussed above, we envisage a simple REG algorithm—hereby called *PersonalityREG* - that uses both context and personality information to perform REG content selection. This is illustrated in Algorithm 1.

Algorithm 1 A REG strategy based on pre-trained classifiers.

```
1: function MakeDescription(r, C, b)
2:     L ← {}
3:     A ← attributes(r)
4:     for a_i ∈ A do
5:         v ← value(r, a_i)
6:         if Predicts[a_i, r, C, b] == True then
7:             L ← L ∪ (a_i, v)
8:     return L
```

The algorithm takes as an input the target r to be described within a context C containing a number of distractor objects and their properties represented as attribute-value pairs and a target personality profile b. As an output, the algorithm returns a set of properties L representing a description of r that resembles what a human speaker with a personality profile b would produce in the same situation.

The above pseudo-code makes use of three auxiliary functions: *attributes*(), which is meant to return the set of all possible attributes of an object (e.g. *gender*, *hair.colour*), *value*(), which returns the value of a given attribute for the referred object (e.g. the value 'blonde' for the attribute *hair.colour*), and *Predicts*(), which invokes the relevant classifier and estimates whether the given attribute would be selected by a human with personality *b* or not in this situation.

The algorithm starts by making an empty output description *L* (line 2) and by obtaining the set of attributes of the target (line 3). For every attribute, the corresponding classifier is invoked (6) and if predicted, the content is selected for inclusion in *L* (7).

To illustrate the kinds of description obtained by this simple procedure, let us consider the goal of generating a uniquely identifying description for the target *o*4 in the previous Fig. 3. Assuming that the pre-trained classifiers would predict the selection of, e.g. *gender*, *race* and *smile* only, the resulting description would take the form of a set of properties *L* = <*gender*-male, *race*-black, *smile*-no>, which could be later realised as, e.g. 'the black man who looks serious'.

As in [32] and similar machine learning approaches to REG, we notice that the present strategy does not explicitly check whether a given attribute is discriminatory. As a result, the algorithm may end up including a certain amount of redundancy in the output description. This was indeed the case of the previous example, in which the reference to *smile* is redundant, and a shorter description (e.g. 'the black man') would suffice for disambiguation. Although our model does not disregard discriminatory power entirely (i.e. discriminatory power values are implicitly modelled as learning features), this behaviour contrasts purely algorithmic solutions [30] in which only discriminatory attributes are to be selected. Allowing a certain amount of redundant information is however common in human language production [51], and allowing the selection of non-discriminatory attributes may be crucial in some (or perhaps most) domains, as illustrated by the case of *isYoung* and *eyebrows* discussed in the "Data" section.

Evaluation

In order to assess the predictions made by *PersonalityREG*, we make use of three baseline systems as follows. The first system is a straightforward implementation of the incremental approach in [30] that iterates over a pre-defined list of preferred attributes computed from the training data. Despite its popularity in the REG field, however, we notice that this baseline does not have access to the same information provided to the alternatives under consideration and therefore cannot be expected to outperform them. Thus, the incremental baseline is included in the present evaluation for illustration purposes only.

The second baseline system is a simplified version of *PersonalityREG* that does not make use of personality information, and it is therefore similar to the machine learning version of the Dale & Reiter incremental algorithm [30] discussed in the "Content selection for referring expression generation" section. This model, hereby called *ContextREG*, relies exclusively on context features provided by the input scene, and it is intended to investigate the possible benefits of taking personality information into account. As in the incremental approach, this strategy makes us of a list of preferred attributes computed from the training data.

Finally, a third baseline system replaces the personality information in *PersonalityREG* for a unique identifier of each speaker, which is a popular strategy in speaker-dependent

REG (cf. the "Content selection for referring expression generation" section), and not unlike [31] and others. This model, hereby called *IndividualREG*, is intended to investigate whether personality-dependent REG may outperform a highly personalised REG strategy of this kind.

In all machine learning models, individual classifiers for each referential attribute under consideration were built using Decision Tree induction. Given the small size and sparsity of our training dataset, class imbalance was once again minimised by making use of SMOTE [44] oversampling with k=5 neighbours. The three main models—*PersonalityREG*, *ContextREG* and *IndividualREG*—give rise to the following research hypotheses.

> *h2a*: The use of personality information about a target speaker enables a REG model to select referential contents that resemble more closely the choices made by humans in this task if compared to a similar model that does not have access to personality information.
>
> *h2b* The use of personality information about a target speaker enables a REG model to select referential contents that resemble more closely the choices made by humans in this task if compared to a similar model in which personality information has been replaced for the explicit identifier of the target speaker.

Hypothesis *h2a* will be tested by comparing referring expressions (represented as sets of semantic properties) generated by *PersonalityREG* with the same descriptions generated by *ContextREG*. We expect that, on average, descriptions produced by *PersonalityREG* will resemble human descriptions more closely than those generated by *ContextREG*. Hypothesis *h2b* will be tested by comparing descriptions generated by *PersonalityREG* with the same descriptions generated by *individualREG*. We expect that, on average, descriptions produced by *PersonalityREG* will resemble human descriptions more closely than those generated by *IndividualREG*.

Evaluation proper consists of generating every description found in the test data using each of the three models separately and by comparing their output to the original (human-produced) descriptions. As in the case of discourse-level CS, we once again measure the degree of overlap between system and human descriptions by computing Dice coefficients [45].

Classification results

Before discussing the actual CS results, we start by assessing the predictions made by the individual classifiers using 10-fold cross validation over the entire dataset. Results obtained by using each of the three feature sets—context features only, speaker's identifiers, and personality information, respectively—are shown in Table 6.

Generally speaking, the personality-aware classifiers appear to outperform the alternatives. This effect will be made more explicit when these classifiers are put to use as part of the actual REG task discussed in the next section.

Content selection results

For each of the four REG models under evaluation—the *ContextREG*, *IndividualREG*, and Incremental baselines, and the proposed *PersonalityREG* model—Table 7 shows mean Dice scores obtained in the generation of test descriptions referring to each input scene in the corpus.

Table 6 Precision (P), recall (R), and F1-measure (F) 10-fold cross validation results on training data

	Context info			Speaker ids			Personality info		
Class	P	R	F	P	R	F	P	R	F
isYoung	0.56	0.50	0.52	0.34	0.60	0.48	0.50	0.66	**0.58**
Race	0.78	0.81	**0.79**	0.40	0.58	0.49	0.63	0.62	0.65
Emotion	0.68	0.57	0.62	0.55	0.82	0.64	0.70	0.83	**0.73**
Smile	0.70	0.65	**0.69**	0.38	0.64	0.50	0.58	0.65	0.60
Eyebrows	0.62	0.68	0.59	0.61	0.85	0.69	0.74	0.89	0.79
hair.colour	0.70	0.59	0.65	0.47	0.69	0.57	0.65	0.72	**0.66**
hair.length	0.69	0.79	**0.72**	0.52	0.78	0.59	0.66	0.75	0.67
Ponytail	0.86	0.98	0.87	0.81	0.89	0.83	0.89	0.91	**0.89**
eye.colour	0.68	0.76	0.72	0.60	0.82	0.67	0.78	0.89	**0.81**
Mean	0.70	0.70	0.69	0.52	0.74	0.61	0.68	0.77	**0.71**

Best F1 results are highlighted

Regarding hypothesis *h2a* (the use of personality information in REG), we notice that, on average, *PersonalityREG* outperforms its personality-free counterpart *ContextREG*. The difference is significant according to a Wilcoxon signed rank test (W =13618, z = 7.89, p <0.0001). The use of personality information about a target speaker enables the REG model to generate descriptions that resemble more closely those produced by humans. This offers support to hypothesis *h2a*.

Regarding hypothesis *h2b* (the use of personality information versus speaker's identifiers), we notice that, on average, *PersonalityREG* outperforms the speaker-specific strategy *IndividualREG*. The difference is also significant (W =6104, z =4.08, p <0.0001). The use of personality information about a target speaker for content selection is superior the use of speaker's identifiers. This offers support to *h2b*.

Final remarks

This paper has focused on the computer side of human-computer natural language interaction, addressing the issue of how the use of personality information may help the development of more natural or human-like systems of this kind. The present study is

Table 7 Mean Dice coefficients and standard deviation per input scene

	ContextREG		IndividualREG		Incremental		PersonalityREG	
Scene	Mean	SD	Mean	SD	Mean	SD	Mean	SD
01	0.65	0.10	**0.78**	0.16	0.62	0.33	0.77	0.17
02	0.51	0.18	0.56	0.29	0.56	0.29	**0.70**	0.23
03	0.56	0.17	0.62	0.17	0.53	0.16	**0.66**	0.24
04	0.56	0.18	0.58	0.18	0.56	0.13	**0.65**	0.22
05	0.53	0.14	0.54	0.22	0.40	0.12	**0.66**	0.16
06	0.62	0.20	0.63	0.23	0.34	0.26	**0.67**	0.21
07	0.55	0.16	0.62	0.15	0.39	0.15	**0.67**	0.23
08	0.60	0.16	0.70	0.25	**0.79**	0.21	0.73	0.22
09	0.49	0.19	0.67	0.21	0.71	0.27	**0.74**	0.17
10	0.54	0.17	0.61	0.24	0.57	0.32	**0.65**	0.23
11	0.65	0.14	0.62	0.17	**0.71**	0.30	0.63	0.19
12	0.57	0.16	0.59	0.23	0.43	0.19	**0.68**	0.18
Mean	0.57	0.17	0.63	0.21	0.55	0.27	**0.69**	0.20

Best results are highlighted

among the first to establish an (admittedly tentative) relation between personality and content selection (as opposed to the more well-documented relation between personality and surface realisation) and, to the best of our knowledge, is the first of its kind to address this issue at both discourse (or macro planning) and reference (or microplanning) levels.

Using personality information to decide which contents (as opposed to which surface forms) should appear in an output text opens a number of opportunities for customisable NLG. At the discourse level, for instance, storytelling systems may be able to produce narratives in which the very plot is driven by a target personality type, potentially making them more compelling or engaging. Effects of this kind are of course more subtle at the reference level, but personality may still play a similar role by reflecting the preferences of a target audience when describing a particular entity (e.g. by focusing on positive features of a character) Similar applications in education, advertisement and others may also be envisaged.

Despite these opportunities, however, the scale of the present experiments are clearly small, and we are aware that their results represent only a first step towards robust personality-dependent CS. In particular, the present focus on machine learning does not further the issue of which contents may be triggered (or favoured) by certain personality types. An investigation of this kind—which remains currently unsupported by our overly small datasets—is left as future work.

A second aspect of the present study that requires further development is the question of how personality-dependent CS models of the kinds under discussion may actually affect a target user at the receiver end of an NLG system. Once again, previous work in the field has shown that effects of this kind hold for personality-dependent surface realisation and other NLG tasks, but it remains unclear to which extent personality-dependent CS may have a similar effect. A study of this kind is also left as future work.

Acknowledgements
The authors are grateful to the participants in the data collection task and to the b5 corpus team.

Authors' contributions
The first author is mainly responsible for experiment 1, and the second author is mainly responsible for experiment 2. The third author has mainly organised the article, whose revision was carried out by the three authors equally. The author(s) read and approved the final manuscript.

Funding
This work has been partially supported by FAPESP grant # 2016/14223-0 and by the University of São Paulo grant # 668/2018.

Availability of data and materials
The present work is entirely based on the b5 corpus described in [34]. The corpus is publicly available for reuse, and it may be downloaded from the link provided in the original publication.

Competing interests
The authors declare that they have no competing interests.

Received: 23 February 2020 Accepted: 17 April 2020
Published online: 29 April 2020

References

1. Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, Sykes C (2009) Automatic generation of textual summaries from neonatal intensive care data. Artif Intell 173:789–816
2. di Eugenio B, Boyd A, Lugaresi C, Balasubramanian A, Keenan G, Burton M, Macieira TGR, Li J, Lussier Y (2014) PatientNarr: Towards generating patient-centric summaries of hospital stays. In: Proceedings of the 8th International Natural Language Generation Conference (INLG-2014). Association for Computational Linguistics, Philadelphia. pp 6–10

3. Jordan P, Green N, Thomas C, Holm S (2014) TBI-Doc: Generating patient & clinician reports from brain imaging data. In: Proceedings of the 8th International Natural Language Generation Conference (INLG)-2014. Association for Computational Linguistics, Philadelphia. pp 143–146
4. Hunter J, Freer Y, Gatt A, Reiter E, Sripada S, Sykes C (2012) Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med 56(3):157–172
5. Reiter E, Robertson R, Osman L (2003) Lessons from a failure: Generating tailored smoking cessation letters. Artif Intell 144:41–58
6. Reiter E, Sripada S, Hunter J, Yu J (2005) Choosing words in computer-generated weather forecasts. Artif Intell 167:137–169
7. Walker MA, Grant R, Sawyer J, Lin GI, Wardrip-Fruin N, Buell M (2011) Perceived or not perceived: Film character models for expressive NLG. In: Si M, Thue D, André E, Lester JC, Tanenbaum J, Zammitto V (eds). Lecture Notes in Computer Science 7069. Springer, Vancouver. pp 109–121
8. Walker M, Lin G, Sawyer J, Grant R, Buell M, Wardrip-Fruin N (2011) Murder in the arboretum: Comparing character models to personality models. In: Intelligent Narrative Technologies. AAAI, Santa Cruz
9. Zhang X, Lapata M (2014) Chinese poetry generation with recurrent neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha. pp 670–680
10. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Boston. pp 3128–3137. https://doi.org/10.1109/CVPR.2015.7298932
11. Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015. PMLR, Lille. pp 2048–2057
12. Reiter E, Dale R (2000) Building Natural Language Generation Systems. Cambridge University Press, New York
13. Goldberg LR (1990) An alternative description of personality: The Big-Five factor structure. J Pers Soc Psychol 59:1216–1229
14. Plank B, Hovy D (2015) Personality traits on Twitter - or - how to get 1,500 personality tests in a week. In: Proc. of WASSA-2015. Association for Computational Linguistics, Lisbon. pp 92–98
15. Álvarez-Carmona M, López-Monroy A, Montes-y-Gómez M, Villaseñor-Pineda L, Escalante H (2015) INAOE's participation at PAN'15: Author Profiling task. In: CLEF 2015. CEUR-WS.org, Toulouse
16. González-Gallardo C, et al. (2015) Tweets classification using corpus dependent tags, character and POS N-grams. In: CLEF 2015. CEUR-WS.org, Toulouse
17. Mairesse F, Walker M (2011) Controlling user perceptions of linguistic style: Trainable generation of personality traits. Comput Linguist 37(3):455–488
18. Herzig J, Shmueli-Scheuer M, Sandbank T, Konopnicki D (2017) Neural response generation for customer service based on personality traits. In: Proceedings of the 10th International Conference on Natural Language Generation. INLG, Santiago de Compostela. pp 252–256
19. Mairesse F, Walker M (2010) Towards personality-based user adaptation: psychologically informed stylistic language generation. User Modeling and User-Adapted Interaction 20(3):227–278
20. Mairesse F, Walker M (2007) PERSONAGE: personality generation for dialogue. In: 45th Annual Meeting-Association For Computational Linguistics. Association for Computational Linguistics (ACL), Sheffield. pp 496–503
21. Paraboni I, van Deemter K (1999) Issues for the generation of document deixis. In: Procs. of Workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in Association with the 11th European Summers School in Logic, Language and Information (esslli99). European Summer School for Language, Logic and Information, Utrecht. pp 44–48
22. Khosmood F, Walker M (2010) Grapevine: a gossip generation system. In: Proceedings of the Fifth International Conference on the Foundations of Digital Games. ACM, Monterey. pp 92–99
23. McCormick C (2012) Evaluating the perception of personality and naturalness in computer generated utterances. In: Master of Sciences dissertation. University of Dublin, Trinity College, Dublin
24. Lukin SM, Ryan JO, Walker M (2014) Automating direct speech variations in stories and games. In: Proceedings of the 3rd Workshop on Games and NLP (GAMNLP 2014). North Carolina State University, Raleigh. pp 1–6. Association for the Advancement of Artificial Intelligence (www.aaai.org). Accessed 3 Oct 2014
25. Bowden KK, Lin G, Reed L, Tree JEF, Walker MA (2016) M2d: monolog to dialog generation for conversational story telling. In: International Conference on Interactive Digital Storytelling. Springer, Switzerland. pp 12–24
26. Aly A, Tapus A (2016) Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. Auton Robots 40(2):193–209
27. dos Santos Silva D, Paraboni I (2015) Generating spatial referring expressions in interactive 3D worlds. Spat Cogn Comput 15(03):186–225. https://doi.org/10.1080/13875868.2015.1039166
28. Krahmer E, van Deemter K (2012) Computational generation of referring expressions: A survey. Comput Linguist 38(1):173–218
29. van Deemter K (2016) Computational Models of Referring. A Study in Cognitive Science. MIT Press, Cambridge
30. Dale R, Reiter E (1995) Computational interpretations of the Gricean maxims in the generation of referring expressions. Cogn Sci 19. https://doi.org/10.1207/s15516709cog1902_3
31. Viethen J, Dale R (2010) Speaker-dependent variation in content selection for referring expression generation. In: Proceedings of the Australasian Language Technology Association Workshop 2010. Australasian Language Technology Association, Melbourne. pp 81–89
32. Ferreira TC, Paraboni I (2014) Referring expression generation: taking speakers' preferences into account. In: Text, Speech and Dialogue (TSD-2014), Lecture Notes in Artificial Intelligence 8655. Springer, Brno. pp 539–546
33. Paraboni I, Monteiro DS, Lan AGJ (2017) Personality-dependent referring expression generation. In: Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence Vol. 10415. Springer, Prague. pp 20–28

34. Ramos RMS, Neto GBS, Silva BBC, Monteiro DS, Paraboni I, Dias RFS (2018) Building a corpus for personality-dependent natural language understanding and generation. In: 11th International Conference on Language Resources and Evaluation (LREC-2018). ELRA, Miyazaki. pp 1138–1145
35. John OP, Donahue E, Kentle R (1991) The Big Five inventory - versions 4a and 54. Technical report, Inst. Personality Social Research, University of California, Berkeley, CA, USA
36. dos Santos VG, Paraboni I, Silva BBC (2017) Big five personality recognition from multiple text genres. In: Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence Vol. 10415. Springer, Prague, Czech Republic. pp 29–37. https://doi.org/10.1007/978-3-319-64206-2_4
37. Silva BBC, Paraboni I (2018) Learning personality traits from Facebook text. IEEE Latin Am Trans 16(4):1256–1262. https://doi.org/10.1109/TLA.2018.8362165
38. Silva BBC, Paraboni I (2018) Personality recognition from Facebook text. In: 13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS Vol. 11122. Springer, Canela. pp 107–114. https://doi.org/10.1007/978-3-319-99722-3_11
39. dos Santos WR, Ramos RMS, Paraboni I (2020) Computational personality recognition from facebook text: psycholinguistic features, words and facets. New Rev Hypermedia Multimed 25(4):268–287. https://doi.org/10.1080/13614568.2020.1722761
40. Hsieh FC, Dias RFS, Paraboni I (2018) Author profiling from Facebook corpora. In: 11th International Conference on Language Resources and Evaluation (LREC-2018). ELRA, Miyazaki. pp 2566–2570
41. Dan-Glauser ES, Scherer KR (2011) The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. Behav Res Methods 43(2):468–477. https://doi.org/10.3758/s13428-011-0064-1
42. Bick E (2000) The parsing system Palavras - automatic grammatical analysis of portuguese in a constraint grammar famework. PhD thesis, Aarhus University
43. Ferreira TC, Paraboni I (2014) Classification-based referring expression generation. In: Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403. Springer, Kathmandu. pp 481–491
44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Int Res 16(1):321–357
45. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302
46. Gatt A, van der Sluis I, van Deemter K (2007) Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: Proceedings of ENLG-07. Association for Computational Linguistics, Schloss Dagstuhl
47. Dale R, Viethen J (2009) Referring expression generation through attribute-based heuristics. In: 12th European Workshop on Natural Language Generation, ENLG '09, Athens. pp 58–65
48. Paraboni I, Galindo M, Iacovelli D (2017) Stars2: a corpus of object descriptions in a visual domain. Lang Resour Eval 51(2):439–462. https://doi.org/10.1007/s10579-016-9350-y
49. Righi G, Peissig JJ, Tarr MJ (2012) Recognizing disguised faces. Vis Cogn 20(2):143–169. https://doi.org/10.1080/13506285.2012.654624
50. Dale R (2002) Cooking up referring expressions. In: Proceedings of ACL-2002. Association for Computational Linguistics, Philadelphia. pp 68–75
51. Pechmann T (1989) Incremental speech production and referential overspecification. Linguistics 27(1):98–110

Publisher's Note