Journal of Artificial Intelligence Research 30 (2007) 457-500 Submitted 05/07; published 11/07

Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text

François Mairesse F.MAIRESSE@SHEFFIELD.AC.UK
Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom

Marilyn A. Walker M.A.WALKER@SHEFFIELD.AC.UK
Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom

Matthias R. Mehl MEHL@EMAIL.ARIZONA.EDU
Department of Psychology, University of Arizona
1503 E University Blvd. Building 68, Tucson, AZ 85721, USA

Roger K. Moore R.K.MOORE@DCS.SHEF.AC.UK
Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom

Abstract

It is well known that utterances convey a great deal of information about the *speaker* in addition to their semantic content. One such type of information consists of cues to the speaker's *personality traits*, the most fundamental dimension of variation between humans. Recent work explores the automatic detection of other types of pragmatic variation in text and conversation, such as emotion, deception, speaker charisma, dominance, point of view, subjectivity, opinion and sentiment. Personality affects these other aspects of linguistic production, and thus personality recognition may be useful for these tasks, in addition to many other potential applications. However, to date, there is little work on the automatic recognition of personality traits. This article reports experimental results for recognition of all Big Five personality traits, in both conversation and text, utilising both self and observer ratings of personality. While other work reports classification results, we experiment with classification, regression and ranking models. For each model, we analyse the effect of different feature sets on accuracy. Results show that for some traits, any type of statistical model performs significantly better than the baseline, but ranking models perform best overall. We also present an experiment suggesting that ranking models are more accurate than multi-class classifiers for modelling personality. In addition, recognition models trained on observed personality perform better than models trained using self-reports, and the optimal feature set depends on the personality trait. A qualitative analysis of the learned models confirms previous findings linking language and personality, while revealing many new linguistic markers.

1. Introduction

> *Personality* is the complex of all the attributes—behavioural, temperamental, emotional and mental—that characterise a unique individual.

It is well known that utterances convey a great deal of information about the *speaker* in addition to their semantic content. One such type of information consists of cues to the speaker's *personality traits*, the most fundamental dimension of variation between humans. Personality is typically assessed along five dimensions known as the Big Five:

- Extraversion vs. Introversion (sociable, assertive, playful vs. aloof, reserved, shy)
- Emotional stability vs. Neuroticism (calm, unemotional vs. insecure, anxious)
- Agreeableness vs. Disagreeable (friendly, cooperative vs. antagonistic, faultfinding)
- Conscientiousness vs. Unconscientious (self-disciplined, organised vs. inefficient, careless)
- Openness to experience (intellectual, insightful vs. shallow, unimaginative)

These five personality traits have been repeatedly obtained by applying factor analyses to various lists of trait adjectives used in personality description questionnaires (sample adjectives above) (Norman, 1963; Peabody & Goldberg, 1989; Goldberg, 1990). The basis for such factor analyses is the *Lexical Hypothesis* (Allport & Odbert, 1936), i.e. that the most relevant individual differences are encoded into the language, and the more important the difference, the more likely it is to be expressed as a single word. Despite some known limits (Eysenck, 1991; Paunonen & Jackson, 2000), over the last 50 years the Big Five model has become a standard in psychology and experiments using the Big Five have shown that personality traits influence many aspects of task-related individual behaviour. For example, the success of most interpersonal tasks depends on the personalities of the participants, and personality traits influence leadership ability (Hogan, Curphy, & Hogan, 1994), general job performance (Furnham, Jackson, & Miller, 1999), attitude toward machines (Sigurdsson, 1991), sales ability (Furnham et al., 1999), teacher effectiveness (Rushton, Murray, & Erdle, 1987), and academic ability and motivation (Furnham & Mitchell, 1991; Komarraju & Karau, 2005). However, to date there has been little work on the automatic recognition of personality traits (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Mairesse & Walker, 2006a, 2006b; Oberlander & Nowson, 2006).

Recent work in AI explores methods for the automatic detection of other types of pragmatic variation in text and conversation, such as emotion (Oudeyer, 2002; Liscombe, Venditti, & Hirschberg, 2003), deception (Newman, Pennebaker, Berry, & Richards, 2003; Enos, Benus, Cautin, Graciarena, Hirschberg, & Shriberg, 2006; Graciarena, Shriberg, Stolcke, Enos, Hirschberg, & Kajarekar, 2006; Hirschberg, Benus, Brenier, Enos, Friedman, Gilman, Girand, Graciarena, Kathol, Michaelis, Pellom, Shriberg, & Stolcke, 2005), speaker charisma (Rosenberg & Hirschberg, 2005), mood (Mishne, 2005), dominance in meetings (Rienks & Heylen, 2006), point of view or subjectivity (Wilson, Wiebe, & Hwa, 2004; Wiebe, Wilson, Bruce, Bell, & Martin, 2004; Wiebe & Riloff, 2005; Stoyanov, Cardie, & Wiebe, 2005; Somasundaran, Ruppenhofer, & Wiebe, 2007), and sentiment or opinion (Turney, 2002; Pang & Lee, 2005; Popescu & Etzioni, 2005; Breck, Choi, & Cardie, 2007). In contrast with these pragmatic phenomena, which may be relatively contextualised or short-lived, personality is usually considered to be a longer term, more stable, aspect of individuals (Scherer, 2003). However, there is evidence that personality interacts with, and affects, these other aspects of linguistic production. For example, there are strong relations between the extraversion and conscientiousness traits and the positive affects, and between neuroticism and disagreeableness and various negative affects (Watson & Clark, 1992). Lying leads to inconsistencies in impressions of the agreeableness personality trait across modes (visual vs. acoustic), and these inconsistencies are used as cues for deception detection by human judges (Heinrich & Borkenau, 1998). Outgoing and energetic people (i.e. extravert) are more successful at deception, while apprehensive (i.e. neurotic) individuals are not as successful (Riggio, Salinas, & Tucker, 1988), and individuals who score highly on the agreeableness and openness to experience traits are also better at detecting deception (Enos et al., 2006). Features used to automatically recognise introversion and extraversion in our studies are also important for automatically identifying deception (Newman et al., 2003). Speaker charisma has been shown to correlate strongly with extraversion (Bono & Judge, 2004), and individuals who dominate meetings have similar characteristics to extraverts, such as verbosity (Rienks & Heylen, 2006). Oberlander and Nowson (2006) suggest that opinion mining could benefit from personality information. Thus this evidence suggests that incorporating personality models into these other tasks may improve accuracy.

We also hypothesise that computational recognition of user personality could be useful in many other computational applications. Identification of leaders using personality dimensions could be useful in analysing meetings and the conversations of suspected terrorists (Hogan et al., 1994; Tucker & Whittaker, 2004; Nunn, 2005). Dating websites could analyse text messages to try to match personalities and increase the chances of a successful relationship (Donnellan, Conger, & Bryant, 2004). Tutoring systems might be more effective if they could adapt to the learner's personality (Komarraju & Karau, 2005). Automatically identifying the author's personality in a corpus could also improve language generation, as individual differences in language affect the way that concepts are expressed (Reiter & Sripada, 2004). Studies have also shown that users' evaluation of conversational agents depends on their own personality (Reeves & Nass, 1996; Cassell & Bickmore, 2003), which suggests a requirement for such systems to adapt to the user's personality, like humans do (Funder & Sneed, 1993; McLarney-Vesotski, Bernieri, & Rempala, 2006).

While in some applications it would be possible to acquire personality information by asking the user or author directly (John, Donahue, & Kentle, 1991; Costa & McCrae, 1992), here we explore whether it is possible to acquire personality models for the Big Five personality traits by observation of individual linguistic outputs in text and conversation. To date, we know of only two studies besides our own on *automatic recognition* of user personality (Argamon et al., 2005; Mairesse & Walker, 2006a, 2006b; Oberlander & Nowson, 2006). Other work has applied classification models to the recognition of personality in texts and blog postings. To our knowledge, the results presented here are the first to examine the recognition of personality in dialogue (Mairesse & Walker, 2006a, 2006b), and to apply regression and ranking models that allow us to model personality recognition using the continuous scales traditional in psychology. We also systematically examine the use of different feature sets, suggested by psycholinguistic research, and report statistically significant results.

We start in Section 2 by reviewing the psychology findings linking personality and language; these findings motivate the features used in the learning experiments described in Section 3. Section 3 overviews the methods we use to automatically train personality models, using both conversation and written language samples, and both self-ratings and observer ratings of personality traits. We explore the use of classification models (Section 4), regression models (Section 5), and ranking models (Section 6), and the effect of different feature sets on model accuracy. The results show that for some traits, any type of statistical model performs significantly better than the baseline, but ranking models perform best overall. In addition, models trained on observed personality scores perform better than models trained using self-reports, and the optimal feature set is dependent on the personality trait. The rules derived and features used in the learned models confirm previous findings linking language and personality, while revealing many new linguistic markers. We delay the review of Argamon et al. (2005) and Oberlander and Nowson (2006) to Section 7, when we can better compare their results with our own, and sum up and discuss future work in Section 8.

2. Personality Markers in Language

Why do we believe it might be possible to automatically recognise personality from linguistic cues? Psychologists have documented the existence of such cues by discovering correlations between a range of linguistic variables and personality traits, across a wide range of linguistic levels, including acoustic parameters (Smith, Brown, Strong, & Rencher, 1975; Scherer, 1979), lexical categories (Pennebaker & King, 1999; Pennebaker, Mehl, & Niederhoffer, 2003; Mehl, Gosling, & Pennebaker, 2006; Fast & Funder, 2007), n-grams (Oberlander & Gill, 2006), and speech-act type (Vogel & Vogel, 1986). As the correlations reported in the literature are generally weak (see Section 3.3), it is not clear whether these features will improve accuracies of statistical models on unseen subjects. Of all Big Five traits, extraversion has received the most attention from researchers. However, studies focusing systematically on all Big Five traits are becoming more common.

2.1 Markers of Extraversion

We summarise various findings linking extraversion and language cues in Table 1, for different levels of language production such as speech, syntax and content selection. A review by Furnham (1990) describes linguistic features linked to extraversion and other traits, and Dewaele and Furnham (1999) review studies focusing on the link between extraversion and both language learning and speech production.

Findings include that there is a higher correlation between extraversion and oral language, especially when the study involves a complex task. Extraverts talk more, louder and more repetitively, with fewer pauses and hesitations, they have higher speech rates, shorter silences, a higher verbal output, a lower type/token ratio and a less formal language, while introverts use a broader vocabulary (Scherer, 1979; Furnham, 1990; Gill & Oberlander, 2002). Extraverts also use more positive emotion words, and show more agreements and compliments than introverts (Pennebaker & King, 1999). Extravert students learning French as a second language produce more back-channels, and have a more implicit style and a lower lexical richness in formal situations. It seems that the more complex the task and the higher the level of anxiety, the easier it is to differentiate between introverts and extraverts (Dewaele & Furnham, 1999).

Heylighen and Dewaele (2002) also note that extraversion is significantly correlated with contextuality, as opposed to formality. Contextuality can be seen a high reliance on shared knowledge between conversational partners, leading to the use of many deictic

Level	Introvert	Extravert
Conversational behaviour	Listen Less back-channel behaviour	Initiate conversation More back-channel behaviour
Topic selection	Self-focused Problem talk, dissatisfaction Strict selection Single topic Few semantic errors Few self-references	Not self-focused* Pleasure talk, agreement, compliment Think out loud* Many topics Many semantic errors Many self-references
Style	Formal Many hedges (tentative words)	Informal Few hedges (tentative words)
Syntax	Many nouns, adjectives, prepositions (explicit) Elaborated constructions Many words per sentence Many articles Many negations	Many verbs, adverbs, pronouns (implicit) Simple constructions* Few words per sentence Few articles Few negations
Lexicon	Correct Rich High diversity Many exclusive and inclusive words Few social words Few positive emotion words Many negative emotion words	Loose* Poor Low diversity Few exclusive and inclusive words Many social words Many positive emotion words Few negative emotion words
Speech	Received accent Slow speech rate Few disfluencies Many unfilled pauses Long response latency Quiet Low voice quality Non-nasal voice Low frequency variability	Local accent* High speech rate Many disfluencies* Few unfilled pauses Short response latency Loud High voice quality Nasal voice High frequency variability

Table 1: Summary of identified language cues for extraversion and various production levels, based on previous studies by Scherer (1979), Furnham (1990), Pennebaker and King (1999), Dewaele and Furnham (1999), Gill (2003), Mehl et al. (2006). Asterisks indicate that the cue is only based on a hypothesis, as opposed to study results.

expressions such as pronouns, verbs, adverbs and interjections, whereas formal language is less ambiguous and assumes less common knowledge. In order to measure this variation, Heylighen and Dewaele suggest the use of a metric called formality, defined as:

$$F = (\text{noun freq} + \text{adjective freq} + \text{preposition freq} + \text{article freq} - \text{pronoun freq} - \text{verb freq} - \text{adverb freq} - \text{interjection freq} + 100)/2$$

They argue that this measure is the most important dimension of variation between linguistic expressions, as shown in Biber's factor analysis of various genres (Biber, 1988). In addition to introversion, the authors also find that formality correlates positively with the level of education and the femininity of the speaker. Situational variables related to the use of formal language are the audience size, the time span between dialogues, the unavailability of feedback, difference of backgrounds and spatial location between speakers, as well as the preceding amount of conversation.

Scherer (1979) shows that extraverts are perceived as talking louder and with a more nasal voice, and that American extraverts tend to make fewer pauses, while German extraverts produce more pauses than introverts. Thus personality markers are culture-dependent, even among western societies.

Oberlander and Gill (2006) use content analysis tools and n-gram language models to identify markers in extravert and introvert emails. They replicate previous findings and identify new personality markers such as first person singular pronouns (e.g., *I don't*) and formal greetings (e.g., *Hello*) for introversion, while less formal phrases such as *Take care* and *Hi* characterise extraverts.

2.2 Markers of Other Big Five Traits

Pennebaker and King (1999) identify many linguistic features associated with each of the Big Five personality traits. They use their Linguistic Inquiry and Word Count (LIWC) tool to count word categories of essays written by students whose personality has been assessed using a questionnaire. The authors find small but significant correlations between their linguistic dimensions and personality traits. Neurotics use more 1^{st} person singular pronouns, more negative emotion words and less positive emotion words. On the other hand, agreeable people express more positive and fewer negative emotions. They also use fewer articles. Conscientious people avoid negations, negative emotion words and words reflecting discrepancies (e.g., *should* and *would*). Finally, openness to experience is characterised by a preference for longer words and words expressing tentativity (e.g., *perhaps* and *maybe*), as well as the avoidance of 1^{st} person singular pronouns and present tense forms.

Additionally, Mehl et al. (2006) study markers of personality as perceived by observers. They find that the use of words related to insight and the avoidance of past tense indicates openness to experience, and swearing marks disagreeableness. The same authors also show that some linguistic cues vary greatly across gender. For example, males perceived as conscientious produce more filler words, while females don't. Gender differences are also found in markers of self-assessed personality: the use of 2^{nd} person pronouns indicates a conscientious male, but an unconscientious female.

Gill and Oberlander (2003) study correlates of emotional stability: they find that neurotics use more concrete and frequent words. However, they also show that observers don't use those cues correctly, as observer reports of neuroticism correlate negatively with self-reports.

Concerning prosody, Smith et al. (1975) also show that speech rate is positively correlated with perceived competence (conscientiousness), and that speech rate has an inverted-U relationship with benevolence (agreeableness), suggesting a need for non-linear models.

Some traits have produced more findings than others. A reason for this might be that some are more reflected through language, like extraversion. However, it is possible that this focus is a consequence of extraversion being correlated with linguistic cues that can be analysed more easily (e.g., verbosity).

3. Experimental Method

We conduct a set of experiments to examine whether automatically trained models can be used to *recognise* the personality of unseen subjects. Our approach can be summarised in five steps:

1. Collect individual corpora;
2. Collect associated personality ratings for each participant;
3. Extract relevant features from the texts;
4. Build statistical models of the personality ratings based on the features;
5. Test the learned models on the linguistic outputs of unseen individuals.

The following sections describe each of these steps in more detail.

3.1 Sources of Language and Personality

Introvert	**Extravert**
I've been waking up on time so far. What has it been, 5 days? Dear me, I'll never keep it up, being such not a morning person and all. But maybe I'll adjust, or not. I want internet access in my room, I don't have it yet, but I will on Wed??? I think. But that ain't soon enough, cause I got calculus homework [...]	I have some really random thoughts. I want the best things out of life. But I fear that I want too much! What if I fall flat on my face and don't amount to anything. But I feel like I was born to do BIG things on this earth. But who knows... There is this Persian party today.
Neurotic	**Emotionally stable**
One of my friends just barged in, and I jumped in my seat. This is crazy. I should tell him not to do that again. I'm not that fastidious actually. But certain things annoy me. The things that would annoy me would actually annoy any normal human being, so I know I'm not a freak.	I should excel in this sport because I know how to push my body harder than anyone I know, no matter what the test I always push my body harder than everyone else. I want to be the best no matter what the sport or event. I should also be good at this because I love to ride my bike.

Table 2: Extracts from the essays corpus, for participants rated as extremely introvert, extravert, neurotic, and emotionally stable.

We use the data from Pennebaker and King (1999) and Mehl et al. (2006) in our experiments. The first corpus contains 2,479 essays from psychology students (1.9 million words), who were told to write whatever comes into their mind for 20 minutes. The data was collected and analysed by Pennebaker and King (1999); a sample is shown in Table 2.

Introvert	Extravert
- Yeah you would do kilograms. Yeah I see what you're saying. - On Tuesday I have class. I don't know. - I don't know. A16. Yeah, that is kind of cool. - I don't know. I just can't wait to be with you and not have to do this every night, you know? - Yeah. You don't know. Is there a bed in there? Well ok just...	- That's my first yogurt experience here. Really watery. Why? - Damn. New game. - Oh. - That's so rude. That. - Yeah, but he, they like each other. He likes her. - They are going to end up breaking up and he's going to be like.
Unconscientious	**Conscientious**
- With the Chinese. Get it together. - I tried to yell at you through the window. Oh. xxxx's fucking a dumb ass. Look at him. Look at him, dude. Look at him. I wish we had a camera. He's fucking brushing his t-shirt with a tooth brush. Get a kick of it. Don't steal nothing.	- I don't, I don't know for a fact but I would imagine that historically women who have entered prostitution have done so, not everyone, but for the majority out of extreme desperation and I think. I don't know, i think people understand that desperation and they don't don't see [...]

Table 3: Extracts from the EAR corpus, for participants rated as extremely introvert, extravert, unconscientious, and conscientious. Only the participants' utterances are shown.

Personality was assessed by asking each student to fill in the Big Five Inventory questionnaire (John et al., 1991), which asks participants to evaluate on a 5 point scale how well their personality matches a series of descriptions.

The second source of data consists of conversation extracts recorded using an Electronically Activated Recorder (EAR) (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001), collected by Mehl et al. (2006). To preserve the participants' privacy, only random snippets of conversation were recorded. This corpus is much smaller than the essays corpus (96 participants for a total of 97,468 words and 15,269 utterances). While the essays corpus consists only of texts, the EAR corpus contains both sound extracts and transcripts. This corpus therefore allows us to build models of personality recognition from speech. Only the *participants'* utterances were transcribed (not those of their conversational partners), making it impossible to reconstruct whole conversations. Nevertheless, the conversation extracts are less formal than the essays, and personality may be best observed in the absence of behavioural constraints. Table 4 shows that while the essays corpus is much larger than the EAR corpus, the amount of data per subject is comparable, i.e. 766 words per subject for the essays and 1,015 for the EAR corpus. Table 3 shows examples of conversations from the EAR corpus for different personality traits.

For personality ratings, the EAR corpus contains both self-reports and ratings from 18 independent observers. Psychologists use self-reports to facilitate evaluating the personality of a large number of participants, and there are a large number of standard self-report tests. Observers were asked to make their judgments by rating descriptions of the Big Five Inventory (John & Srivastava, 1999) on a 7 point scale (from *strongly disagree* to *strongly*

Dataset	**Essays**	**EAR**
Source of language	Written	Spoken
Personality reports	Self reports	Self and observer
Number of words	1.9 million	97,468
Subjects	2,479	96
Words per subject	766.4	1,015.3

Table 4: Comparison of the essays and EAR corpora.

agree), without knowing the participants. Observers were divided into three groups, each rating one third of the participants, after listening to each participant's entire set of sound files (130 files on average). The personality assessment was based on the audio recordings, which contain more information than the transcripts (e.g., ambient sounds, including captured conversations). Mehl et al. (2006) report strong inter-observer reliabilities across all Big Five dimensions (intraclass correlations based on one-way random effect models: mean $r = 0.84$, $p < .01$). The observers' ratings were averaged for each participant, to produce the final scores used in our experiments.

Interestingly, the average correlations between frequency counts from psycholinguistic word categories and the Big Five personality dimensions were considerably larger in the EAR corpus than with the student essays studied by Pennebaker and King. Moreover, the correlations reported by Mehl et al. seem to be higher for observer reports than for self-reports. Based on this observation, we hypothesise that models of observed personality will outperform models of self-assessed personality.

3.2 Features

The features used in the experiments are motivated by previous psychological findings about correlations between measurable linguistic factors and personality traits. Features are divided into subsets depending on their source and described in the subsections below. The total feature set is summarised in Table 6. The experimental results given in Sections 4, 5, and 6 examine the effect of each feature subset on model accuracy.

3.2.1 Content and Syntax

We extracted a set of linguistic features from each essay and conversation transcript, starting with frequency counts of 88 word categories from the Linguistic Inquiry and Word Count (LIWC) utility (Pennebaker et al., 2001). These features include both syntactic (e.g., ratio of pronouns) and semantic information (e.g., positive emotion words), which were validated by expert judges. Some LIWC features are illustrated in Table 5. Pennebaker and King (1999) previously found significant correlations between these features and each of the Big Five personality traits. Relevant word categories for extraversion include social words, emotion words, first person pronouns, and present tense verbs. Mehl et al. (2006) showed that LIWC features extracted from the EAR corpus were significantly correlated with both self and observer reports of personality.

We also added 14 additional features from the MRC Psycholinguistic database (Coltheart, 1981), which contains statistics for over 150,000 words, such as estimates of the age

Feature	Type	Example
Anger words	LIWC	hate, kill, pissed
Metaphysical issues	LIWC	God, heaven, coffin
Physical state/function	LIWC	ache, breast, sleep
Inclusive words	LIWC	with, and, include
Social processes	LIWC	talk, us, friend
Family members	LIWC	mom, brother, cousin
Past tense verbs	LIWC	walked, were, had
References to friends	LIWC	pal, buddy, coworker
Imagery of words	MRC	Low: future, peace - High: table, car
Syllables per word	MRC	Low: a - High: uncompromisingly
Concreteness	MRC	Low: patience, candor - High: ship
Frequency of use	MRC	Low: duly, nudity - High: he, the

Table 5: Examples of LIWC word categories and MRC psycholinguistic features (Pennebaker et al., 2001; Coltheart, 1981). MRC features associate each word to a numerical value.

of acquisition, frequency of use, and familiarity. As introverts take longer to reflect on their utterances, Heylighen and Dewaele (2002) suggest that their vocabulary is richer and more precise, implying a lower frequency of use. The MRC feature set was previously used by Gill and Oberlander (2002), who showed that extraversion is negatively correlated with concreteness. Concreteness also indicates neuroticism, as well as the use of more frequent words (Gill & Oberlander, 2003). Table 5 shows examples of MRC scales. Each MRC feature is computed by averaging the feature value of all the words in the essay or conversational extract. Part-of-Speech tags are computed to identify the correct entry in the database among a set of homonyms.

3.2.2 Utterance Type

Various facets of personality traits seem to depend on the level of initiative of the speaker and the type of utterance used (e.g., assertiveness, argumentativeness, inquisitiveness, etc.). For example, extraverts are more assertive in their emails (Gill & Oberlander, 2002), while extravert second language learners were shown to produce more back-channel behaviour (Vogel & Vogel, 1986). We therefore introduced features characterising the types of utterance produced. We automatically tagged each utterance of the EAR corpus with speech act categories from Walker and Whittaker (1990), using heuristic rules based on each utterance's parse tree:

- Command: utterance using the imperative form, a command verb (e.g., *must* and *have to*) or a yes/no second person question with a modal auxiliary like *can*;
- Prompt: single word utterance used for back-channelling (e.g., *Yeah*, *OK*, *Huh*, etc.);
- Question: interrogative utterance which isn't a command;
- Assertion: any other utterance.

LIWC FEATURES (Pennebaker et al., 2001): · **Standard counts:** - Word count (WC), words per sentence (WPS), type/token ratio (Unique), words captured (Dic), words longer than 6 letters (Sixltr), negations (Negate), assents (Assent), articles (Article), prepositions (Preps), numbers (Number) - Pronouns (Pronoun): 1^{st} person singular (I), 1^{st} person plural (We), total 1^{st} person (Self), total 2^{nd} person (You), total 3rd person (Other) · **Psychological processes:** - Affective or emotional processes (Affect): positive emotions (Posemo), positive feelings (Posfeel), optimism and energy (Optim), negative emotions (Negemo), anxiety or fear (Anx), anger (Anger), sadness (Sad) - Cognitive Processes (Cogmech): causation (Cause), insight (Insight), discrepancy (Discrep), inhibition (Inhib), tentative (Tentat), certainty (Certain) - Sensory and perceptual processes (Senses): seeing (See), hearing (Hear), feeling (Feel) - Social processes (Social): communication (Comm), other references to people (Othref), friends (Friends), family (Family), humans (Humans) · **Relativity:** - Time (Time), past tense verb (Past), present tense verb (Present), future tense verb (Future) - Space (Space): up (Up), down (Down), inclusive (Incl), exclusive (Excl) - Motion (Motion) · **Personal concerns:** - Occupation (Occup): school (School), work and job (Job), achievement (Achieve) - Leisure activity (Leisure): home (Home), sports (Sports), television and movies (TV), music (Music) - Money and financial issues (Money) - Metaphysical issues (Metaph): religion (Relig), death (Death), physical states and functions (Physcal), body states and symptoms (Body), sexuality (Sexual), eating and drinking (Eating), sleeping (Sleep), Grooming (Groom) · **Other dimensions:** - Punctuation (Allpct): period (Period), comma (Comma), colon (Colon), semi-colon (Semic), question (Qmark), exclamation (Exclam), dash (Dash), quote (Quote), apostrophe (Apostro), parenthesis (Parenth), other (Otherp) - Swear words (Swear), nonfluencies (Nonfl), fillers (Fillers)
MRC FEATURES (Coltheart, 1981): Number of letters (Nlet), phonemes (Nphon), syllables (Nsyl), Kucera-Francis written frequency (K-F-freq), Kucera-Francis number of categories (K-F-ncats), Kucera-Francis number of samples (K-F-nsamp), Thorndike-Lorge written frequency (T-L-freq), Brown verbal frequency (Brown-freq), familiarity rating (Fam), concreteness rating (Conc), imageability rating (Imag), meaningfulness Colorado Norms (Meanc), meaningfulness Paivio Norms (Meanp), age of acquisition (AOA)
UTTERANCE TYPE FEATURES: Ratio of commands (Command), prompts or back-channels (Prompt), questions (Question), assertions (Assertion)
PROSODIC FEATURES: Average, minimum, maximum and standard deviation of the voice's pitch in Hz (Pitch-mean, Pitch-min, Pitch-max, Pitch-stddev) and intensity in dB (Int-mean, Int-min, Int-max, Int-stddev), voiced time (Voiced) and speech rate (Word-per-sec)

Table 6: Description of all features, with feature labels in brackets.

We evaluated the automatic tagger by applying it to a set of 100 hand-labelled utterances randomly selected in the EAR corpus. We obtain 88% of correct labels, which are mostly assertions. Table 7 summarises the partition and the evaluation results for each speech act type. For each speech act, the corresponding feature value is the ratio of the number of occurrences of that speech act to the total number of utterances in each text.

Label	Fraction	Labelling accuracy
Assertion	73.0%	0.95
Command	4.3%	0.50
Prompt	7.0%	0.57
Question	15.7%	1.00
All	100%	0.88

Table 7: Partition of the speech acts automatically extracted from the EAR corpus, and classification accuracies on a sample of 100 hand-labelled utterances.

3.2.3 Prosody

Personality was also shown to influence speech production. Extraversion is associated with more variation of the fundamental frequency (Scherer, 1979), with a higher voice quality and intensity (Mallory & Miller, 1958), and with fewer and shorter silent pauses (Siegman & Pope, 1965). Smith et al. (1975) showed that speech rate is positively correlated with perceived competence (conscientiousness). Interestingly, the same authors found that speech rate has an inverted-U relationship with benevolence (agreeableness), suggesting a need for non-linear models. See Section 3.4.

We added prosodic features based on the audio data of the EAR conversation extracts. As the EAR recorded the participants at anytime of the day, it was necessary to automatically remove any non-voiced signal. We used Praat (Boersma, 2001) to compute features characterising the voice's pitch and intensity (mean, extremas and standard deviation), and we added an estimate of the speech rate by dividing the number of words by the voiced time. As an important aspect of this work is that all features are extracted without any manual annotation beyond transcription, we didn't filter out utterances from other speakers that may have been captured by the EAR even though it utilised a microphone pointing towards the participant's head. Although advances in speaker recognition techniques might improve the accuracy of prosodic features, we make the assumption that the noise introduced by the surrounding speakers has little effect on our prosodic features, and that it therefore does not affect the performance of the statistical models. This assumption still remains to be tested, as the personality similarity-attraction effect (Byrne & Nelson, 1965) might influence the personality distribution of a participant's conversational partners.

We included all the features mentioned in this section (117) in the models based on the EAR corpus. Models computed using the essays corpus contain only LIWC and MRC features (102), as speech acts are only meaningful in dialogues.

3.3 Correlational Analysis

In order to assess what individual features are important for modelling personality regardless of the model used, we report previous correlational studies for the LIWC features on the same data as well as analyses of the new MRC, utterance type and prosodic features. The LIWC features were already analysed by Mehl et al. (2006) for the EAR dataset, and by

Pennebaker and King (1999) for the essays.[1] Tables 8 to 11 show the features correlating significantly with personality ratings ($p < .05$, correlations above .05 only), combining together results from previous studies and new findings that provide insight into the features likely to influence the personality recognition models in Sections 4.3, 5.3 and 6.3.

The correlation magnitudes in Tables 8 and 9 between LIWC and MRC features and the essays data set show that although extraversion is very well perceived in conversations, it isn't strongly reflected through written language, as the correlation magnitudes for the essays dataset are noticeably low. Table 10 shows that word count (WC) is a very important feature for modelling extraversion in conversation, both for observer reports and self-reports. Interestingly, this marker doesn't hold for written language (see Table 9). Other markers common to observed and self-reported extraversion include the variation of intensity (Int-stddev), the mean intensity (Int-mean), word repetitions (Unique), words with a high concreteness (Conc) and imageability (Imag). See Table 11. On the other hand, words related to anger, affect, swearing, and positive and negative emotions (Posemo and Negemo) are perceived as extravert, but they don't mark self-assessed extraversion in conversations.

Tables 10 and 11 show that for emotional stability, only a few markers hold for both self-reports and observer reports: a high word count and a low mean pitch (Pitch-mean). Surprisingly, observed emotional stability is associated with swearing and anger words, but not the self-assessed ratings. As reported by Mehl et al. (2006), neurotics are expected to produce more self-references (Self and I). Pennebaker and King (1999) show that neurotics' use of self-references is also observed in the essays, as well as the use of words related to negative emotions and anxiety. Table 11 shows that in conversations, self-assessed neurotics tend to have a low and constant voice intensity (Int-mean and Int-stddev), while these markers aren't used by observers at all.

While emotional stability is expressed differently in various datasets, some markers of agreeableness are consistent: words related to swearing (Swear) and anger (Anger) indicate both self-assessed and observed disagreeableness, regardless of the source of language. See Tables 8, 9 and 10. Interestingly, Table 11 shows that agreeable people do more back-channelling (Prompt), suggesting that they tend to listen more to their conversational partners. While observers don't seem to take prosody into account for evaluating agreeableness, Table 11 shows that prosodic cues such as the pitch variation (Pitch-stddev) and the maximum voice intensity (Max-int) indicate self-assessed disagreeableness.

As far as markers of conscientiousness are concerned, Tables 8 to 10 show that they are similar to those of agreeableness, as unconscientious participants also use words related to swearing (Swear), anger (Anger) and negative emotions (Negemo), regardless of the dataset and assessment method. On the other hand, observed conscientiousness is associated with words expressing insight, back-channels (Prompt), longer words (Nphon, Nlet, Nsyl and Sixltr) as well as words that are acquired late by children (AOA), while self-assessed conscientiousness is mostly expressed through positive feelings (Posfeel) in conversations. The avoidance of negative language seems to be the main marker of conscientiousness in essays, as all other features in Table 8 correlate only weakly with the self-reports.

1. Our correlations differ from Pennebaker and King's study because we use additional student essays collected during the following years.

Trait	Extraversion	Emotional stability	Agreeableness	Conscientiousness	Openness to experience
LIWC					
Achieve	.03	.01	-.01	.02	-.07**
Affect	.03	-.07**	-.04	-.06**	.04*
AllPct	-.08**	-.04	-.01	-.04	.10**
Anger	-.03	-.08**	-.16**	-.14**	.06**
Anx	-.01	-.14**	.03	.05*	-.04
Apostro	-.08**	-.04	-.02	-.06**	.05**
Article	-.08**	.11**	-.03	.02	.11**
Assent	.01	.02	.00	-.04	.04*
Body	-.05**	-.04	-.04*	-.04*	.02
Cause	.01	-.03	.00	-.04	-.05*
Certain	.05*	-.01	.03	.04*	.04
Cogmech	-.03	-.02	-.02	-.06**	.02
Comm	-.02	.00	-.01	-.05**	.03
Comma	-.02	.01	-.02	-.01	.10**
Death	-.02	-.04	-.02	-.06**	.05*
Dic	.05*	-.09**	.06**	.06**	-.20**
Excl	-.01	.02	-.02	-.01	.07**
Exclam	.00	-.05*	.06**	.00	-.03
Family	.05*	-.05*	.09**	.04*	-.07**
Feel	-.01	-.09**	.04	.02	-.04*
Fillers	-.04*	.01	-.01	-.03	-.01
Friends	.06**	-.04*	.02	.01	-.12**
Future	-.02	.01	.02	.07**	-.04
Groom	-.02	-.02	.01	.01	-.05**
Hear	-.03	.00	-.01	-.04*	.04*
Home	-.01	-.02	.04*	.06**	-.15**
Humans	.04	-.02	-.03	-.08**	.04
I	.05*	-.15**	.05*	.04	-.14**
Incl	.04*	-.01	.03	.04*	-.03
Inhib	-.03	.02	-.02	-.02	.04*
Insight	-.01	-.01	.00	-.03	.05*
Job	.02	.01	.01	.05**	-.05**
Leisure	-.03	.07**	.03	-.01	-.05**
Metaph	-.01	.01	-.01	-.08**	.08**
Motion	.03	-.01	.05*	.03	-.13**
Music	-.04*	.06**	-.01	-.07**	.10**
Negate	-.08**	-.12**	-.11**	-.07**	.01
Negemo	-.03	-.18**	-.11**	-.11**	.04
Nonfl	-.03	.01	.01	-.05*	.02
Number	-.03	.05*	-.03	-.02	-.06**
Occup	.03	.05*	.04	.09**	-.18**
Optim	.03	.04	.01	.08**	-.07**
Other	.06**	-.01	.03	.01	.01
Othref	.07**	.02	.01	.01	.06**
Parenth	-.06**	.03	-.04*	-.01	.10**
Period	-.05*	-.03	-.01	-.01	.04
Physcal	-.02	-.05*	-.03	-.03	.01
Posemo	.07**	.07**	.05*	.02	.02
Posfeel	.07**	-.01	.03	-.02	.08**
Preps	.00	.06**	.04	.08**	-.04
Present	.00	-.12**	-.01	-.03	-.09**
Pronoun	.07**	-.12**	.04*	.02	-.06**
Qmark	-.06**	-.05*	-.04	-.06**	.08**

Table 8: Pearson's correlation coefficients between LIWC features and personality ratings for the essays dataset, based on the analysis from Pennebaker and King (1999) (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

Trait	Extraversion	Emotional stability	Agreeableness	Conscientiousness	Openness to experience
LIWC (2)					
Quote	-.05*	-.02	-.01	-.03	.09**
Relig	.00	.03	.00	-.06**	.07**
Sad	.00	-.12**	.00	.01	-.01
School	.03	.05**	.06**	.10**	-.20**
See	.00	.09**	.00	-.03	.05**
Self	.07**	-.14**	.06**	.04*	-.14**
Semic	-.03	.02	.02	.00	.05**
Sexual	.07**	-.02	.00	-.04	.09**
Sixltr	-.06**	.06**	-.05*	.02	.10**
Sleep	-.01	-.03	-.02	.03	-.08**
Social	.08**	.00	.02	-.02	.02
Space	-.02	.05*	.03	.01	-.04
Sports	.01	.09**	.02	.00	-.05**
Swear	-.01	.00	-.14**	-.11**	.08**
Tentat	-.06**	-.01	-.03	-.06**	.05*
Time	-.02	.02	.07**	.09**	-.15**
TV	-.04	.04*	-.02	-.04*	.04
Unique	-.05**	.10**	-.04*	-.05*	.09**
Up	.03	.06**	.02	-.01	-.06**
WC	.03	-.06**	.01	.02	.05*
We	.06**	.07**	.04*	.01	.04
WPS	-.01	.02	.02	-.02	.06**
You	-.01	.03	-.06**	-.04*	.11**
MRC					
AOA	-.01	.05*	-.04*	.06**	.11**
Brown-freq	.05*	-.06**	.03	.06**	-.07**
Conc	.02	-.06**	.03	-.01	-.10**
Fam	.08**	-.05*	.08**	.05**	-.17**
Imag	.05*	-.04*	.05*	.00	-.08**
K-F-freq	-.01	.10**	.00	.05*	.07**
K-F-ncats	.06**	-.04*	.08**	.07**	-.12**
K-F-nsamp	.06**	-.01	.03	.05**	-.07**
Meanc	.06**	-.10**	.05**	-.01	-.11**
Meanp	.02	-.02	.05*	.00	-.04*
Nlet	-.09**	.09**	-.03	.00	.15**
Nphon	-.08**	.08**	-.03	.01	.14**
Nsyl	-.07**	.07**	-.02	.04	.13**
T-L-freq	.01	.10**	.01	.06**	.05**

Table 9: Continuation of Table 8, i.e. Pearson's correlation coefficients between LIWC and MRC features and personality ratings for the essays dataset (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

Tables 8 and 9 show that openness to experience is the trait yielding the highest correlations in the essays corpus: articles, second person pronouns (You) and long words (Sixltr) indicate openness, while non-open participants tend to talk about their occupations (Occup, Home and School) and themselves (Self). As far as conversations are concerned, observers use similar cues for openness as with conscientiousness, such as insight words, longer words, back-channels and a high age of acquisition (AOA).

This section shows that features are likely to vary depending on the source of language and the method of assessment of personality. While such analyses can help evaluate the usefulness of individual features, the question of how such features should be combined to predict personality accurately is addressed by the statistical models.

Dataset	Observer reports					Self-reports				
Trait	**Extra**	**Emot**	**Agree**	**Consc**	**Open**	**Extra**	**Emot**	**Agree**	**Consc**	**Open**
LIWC										
Affect	.40**	.13	-.20	-.24*	.00	.05	-.13	-.17	-.19	.13
Anger	.37**	.30**	-.49**	-.56**	-.14	-.02	.07	-.30**	-.30**	.10
Articles	.21*	.32**	.03	-.15	.14	.03	.00	.04	-.09	-.04
Assent	-.29**	-.02	.30**	.24*	.03	-.11	-.05	.19	-.03	.08
Cause	-.13	-.23*	.03	.15	.00	.00	-.09	.07	-.02	-.23*
Cogmech	.04	-.01	.24*	.20*	.23*	.11	.01	.08	.00	-.06
Comm	-.18	-.27**	-.14	.00	-.26*	-.01	-.13	.20*	.12	-.17
Dic	-.07	-.16	-.17	-.05	-.08	.02	-.15	.16	-.01	-.20*
Discrep	.08	-.03	.13	.10	.23*	.10	-.01	.15	.09	-.09
Eating	.25*	.15	-.31**	-.43**	-.11	-.03	-.02	-.10	-.19	-.05
Family	.26*	-.23*	-.12	-.03	-.04	.14	-.02	.26**	.04	-.14
Feel	.21*	.06	.03	-.03	.05	.08	.05	-.08	.02	.02
Female	.29**	-.03	.04	.03	-.17	.24*	.07	.29**	.12	-.22*
Filler	-.01	-.19	.04	.20*	.01	-.05	-.13	.20	.18	-.08
Friend	.14	-.01	-.08	-.13	-.14	.20*	.01	.05	.16	-.11
Hear	-.20	-.23*	-.19	-.07	-.29**	-.04	-.08	.13	.07	-.19
Home	-.02	-.19	.03	.04	.06	.04	-.12	.29**	-.03	-.07
Humans	-.01	.21*	-.01	-.23*	-.12	.07	-.03	-.20	-.06	.01
I	.03	-.41**	-.21*	-.08	-.17	.21*	-.16	.23*	.01	-.08
Inhib	.19	.01	-.22*	-.14	.00	.02	.02	-.18	-.11	-.12
Insight	.04	-.02	.34**	.29**	.32**	-.06	-.10	.03	.01	.05
Metaph	.30**	.07	-.10	-.26*	-.02	.20	.10	-.10	-.09	.03
Money	-.02	.24*	-.13	-.24*	.01	-.08	.01	-.22*	-.06	-.15
Negemo	.36**	.18	-.44**	-.49**	-.11	.03	-.05	-.16	-.25*	.10
Nonfl	-.01	.05	.09	.24*	.06	-.02	.17	-.03	-.02	.17
Other	.09	.02	-.07	-.09	-.17	.02	.04	.05	.05	-.28**
Othref	.00	.05	-.13	-.14	-.22*	.02	.13	.07	.01	-.19
Past	-.19	-.07	-.25*	-.18	-.31**	-.10	-.18	-.05	.05	-.26**
Physcal	.30**	.24*	-.39**	-.47**	-.17	-.07	-.06	-.16	-.27**	.05
Posfeel	.28**	.04	.05	.14	.05	.06	-.14	-.07	.23*	.11
Pronoun	-.02	-.30**	-.23*	-.17	-.28**	.12	-.07	.19	.05	-.21*
Relig	.30**	.06	-.09	-.27**	-.07	.26*	.15	-.06	-.09	.04
Self	.09	-.42**	-.25*	-.13	-.15	.25*	-.17	.18	.02	-.08
Senses	-.04	-.12	-.18	-.15	-.26*	.03	-.10	.12	.03	-.14
Sexual	.24*	.21*	-.49**	-.48**	-.22*	-.05	.04	-.19	-.23*	.04
Sixltr	-.04	-.04	.25*	.30**	.24*	-.20	-.15	-.01	.19	.03
Social	-.04	-.06	-.17	-.15	-.31**	.06	.04	.12	.06	-.21*
Space	.03	.18	-.21*	-.24*	-.07	-.10	.09	-.18	.01	.23*
Sports	.10	.28**	-.15	-.19	-.11	.03	.21*	-.15	-.05	-.03
Swear	.30**	.27**	-.51**	-.61**	-.17	-.08	.06	-.28**	-.29**	.06
Tentat	-.04	.15	.26*	.15	.30**	-.14	.04	.05	.14	.05
Unique	-.6**	-.18	-.03	-.03	-.12	-.32**	-.22*	-.18	-.05	-.03
Up	.06	.04	-.08	-.11	-.05	.06	.07	-.05	.03	.31**
WC	.63**	.28**	.10	.07	.20	.29**	.22*	.18	.03	.06

Table 10: Pearson's correlation coefficients between LIWC features and personality ratings for the EAR dataset, based on the analysis from Mehl et al. (2006) (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

3.4 Statistical Models

Various systems require different levels of granularity for modelling personality: it might be more important to cluster users into large groups as correctly as possible, or the system might need to discriminate between individual users. Depending on the application and the adaptation capabilities of the target system, it is possible to use different types of personality models, depending on whether personality modelling is treated as a classification problem, as

Dataset	Observer reports					Self-reports				
Trait	**Extra**	**Emot**	**Agree**	**Consc**	**Open**	**Extra**	**Emot**	**Agree**	**Consc**	**Open**
Prosody										
Int-max	.42**	.12	.07	-.13	.05	.19	.10	-.25*	-.01	.14
Int-mean	.32**	.20	-.02	-.06	.04	.21*	.22*	-.05	-.16	.03
Int-stddev	.40**	.03	-.08	-.12	-.08	.36**	.28**	.00	-.06	.10
Pitch-max	.28**	.10	.13	.05	.23*	-.03	-.11	-.10	-.03	.01
Pitch-mean	.17	-.45**	.06	.04	-.18	.12	-.25*	.07	.03	-.04
Pitch-min	-.17	-.23*	-.02	.08	-.04	.09	-.08	.21*	.04	.08
Pitch-stddev	-.13	.13	.07	.03	.11	-.28**	.01	-.34**	.03	-.03
Voiced	.23*	.27**	.06	.03	.21*	-.02	.07	-.04	-.03	.03
Word-per-sec	.07	-.14	-.12	-.04	-.17	.20*	.07	.09	.02	.04
MRC										
AOA	-.23*	.01	.26**	.26**	.21*	-.12	.04	.05	-.05	.08
Brown-freq	-.26*	-.41**	-.08	.07	-.16	-.04	-.15	.14	.07	-.12
Conc	.24*	-.05	-.20*	-.33**	-.32**	.23*	-.10	.01	-.12	-.02
Fam	-.17	-.28**	-.24*	-.07	-.18	-.03	-.21*	.17	.01	-.13
Imag	.33**	.00	-.23*	-.33**	-.35**	.25*	-.09	.01	-.06	-.03
K-F-freq	-.27**	-.04	.07	.17	.16	-.22*	-.06	-.24*	.05	-.01
K-F-ncats	-.24*	-.24*	-.03	.08	.00	-.01	-.06	.17	.05	-.23*
K-F-nsamp	-.24*	-.20*	-.03	.16	.20	-.15	-.04	.03	.08	-.17
Meanc	.29**	-.10	-.18	-.25*	-.34**	.23*	-.12	.08	-.06	-.07
Nlet	-.14	.17	.25*	.31**	.25*	-.23*	.03	-.18	.13	.12
Nphon	-.12	.09	.25*	.36**	.28**	-.16	.02	-.20	.15	.13
Nsyl	-.16	-.04	.23*	.34**	.19	-.13	-.02	-.06	.12	.10
T-L-freq	-.24*	-.06	.06	.16	.13	-.19	-.07	-.18	.06	-.08
Utterance type										
Assertion	-.05	-.21*	-.03	.01	-.09	-.02	-.06	-.09	.21*	-.14
Command	.00	.01	-.08	-.20*	.00	.13	.21*	-.01	.00	.16
Prompt	-.10	.07	.36**	.27**	.25*	-.05	.01	.22*	-.05	.02
Question	.13	.22*	-.16	-.11	-.04	.01	-.01	-.02	-.24*	.10

Table 11: Continuation of Table 10, i.e. Pearson's correlation coefficients between features and personality ratings for the EAR dataset (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

in previous work by Argamon et al. (2005) and Oberlander and Nowson (2006), or whether we model personality traits via the scalar values actually generated by the self-reports and observer methods used in the corpus collection described in Section 3.1.

To support applications in dialogue system adaptation, where the output generation is limited to a few points at extremes of a personality scale, such as introvert vs. extravert language or neurotic vs. emotionally stable, we develop classification models by splitting our subjects into two equal size groups.

However, if we model personality traits as scalar values, we have two choices. We can treat personality modelling as a regression problem or as a ranking problem. While regression models can replicate the actual scalar values seen in the personality ratings data, there is also a good argument for treating personality as a ranking problem because by definition, personality evaluation assesses relative differences between individuals, e.g. one person is described as an extravert because the average population is not. Moreover, Freund, Iyer, Schapire, and Singer (1998) argue that ranking models are a better fit to learning problems in which scales have arbitrary values (rather than reflecting real world measures).

For classification and regression models, we use the Weka toolbox (Witten & Frank, 2005) for training and evaluation. In order to evaluate models of personality classification, we compare six different learning algorithms against a baseline returning the majority class. The classification algorithms analysed here are C4.5 decision tree learning (J48), Nearest neighbour ($k = 1$), Naive Bayes (NB), Ripper (JRip), Adaboost (10 rounds of boosting) and Support vector machines with linear kernels (SMO).

For regression, we compare five algorithms with a baseline model returning the mean personality score. We focus on a linear regression model, an M5' regression tree, an M5' model tree returning a linear model, a REPTree decision tree, and a model based on Support vector machines with linear kernels (SMOreg). Parameters of the algorithms are set to Weka's default values.

Concerning the ranking problem, we train personality models for each Big Five trait using RankBoost, a boosting algorithm for ranking (Freund et al., 1998; Schapire, 1999). Given a personality trait to model, the linguistic features and personality scores are converted into a training set $\mathcal{T}$ of *ordered pairs* of examples x, y:

$$\begin{aligned}\mathcal{T} = \{(x,y)|\ & x, y \text{ are language samples from two individuals,}\\ & x \text{ has a higher score than } y \text{ for that personality trait}\}\end{aligned}$$

Each example x is represented by a set of m indicator functions $h_s(x)$ for $1 \leq s \leq m$. The indicator functions are calculated by thresholding the feature values (counts) described in Section 3.2. For example, one indicator function is:

$$h_{100}(x) = \begin{cases} 1 & \text{if } \textsc{Word-per-sec}(x) \geq 0.73 \\ 0 & \text{otherwise} \end{cases}$$

So $h_{100}(x) = 1$ if x's average speech rate is above 0.73 words per second. A single parameter α_s is associated with each indicator function, and the *ranking score* for an example x is calculated as

$$F(x) = \sum_s \alpha_s h_s(x)$$

This score is used to rank various language samples (written text or conversation extracts), with the goal of duplicating the ranking found in the training data, and the training examples are used to set the parameter values α_s. Training is the process of setting the parameters α_s to minimise the following loss function:

$$Loss = \frac{1}{|\mathcal{T}|} \sum_{(x,y)\in\mathcal{T}} eval(F(x) \leq F(y))$$

The *eval* function returns 1 if the ranking scores of the (x, y) pair are misordered, and 0 otherwise. In other words, the ranking loss is the percentage of misordered pairs, for which the order of the predicted scores doesn't match the order dictated by the personality scores from the questionnaire.

Most of the techniques used in this work express the learned models as rules or decision trees, which support the analysis of differences in the personality models (see Sections 4.3, 5.3 and 6.3).

4. Classification Results

We evaluate binary classification models based on the essays corpus with self-reports of personality, as well as models based on the EAR corpus with both self and observer reports. All results are averaged over a 10-fold cross-validation, and all significance tests were done using a two-tailed paired t-test at the $p < .05$ level.

4.1 Essays Corpus

Classification results for the essays corpus with self-reports are in Table 12. Interestingly, openness to experience is the easiest trait to model as five classifiers out of six significantly outperform the baseline and four of them produce their best performance for that trait, with accuracies up to 62.1% using support vector machines (SMO). Emotional stability produces the second best performance for four classifiers out of six, with 57.4% accuracy for the SMO model. Conscientiousness is the hardest trait to model as only two classifiers significantly outperform the baseline, however the SMO model performs as well as the best model for extraversion and agreeableness, with around 55% correct classifications.

We find that support vector machines generally perform the best, with Naive Bayes and AdaboostM1 in second position. SMO significantly outperforms the majority class baseline for each trait. A J48 decision tree for recognising extraversion is shown in Figure 1, and the rule-based JRip model classifying openness to experience with 58.8% accuracy is illustrated in Table 16.

Trait	**Base**	**J48**	**NN**	**NB**	**JRIP**	**ADA**	**SMO**
Extraversion	50.04	54.44•	53.27•	53.35•	52.70	55.00 •	54.93 •
Emotional stability	50.08	51.09	51.62	56.42•	55.90 •	55.98 •	57.35 •
Agreeableness	50.36	53.51•	50.16	53.88•	52.63	52.71	55.78 •
Conscientiousness	50.57	51.37	52.10	53.80	52.71	54.45 •	55.29 •
Openness to experience	50.32	54.24•	53.07	59.57•	58.85 •	59.09 •	62.11 •

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 12: Classification accuracy with two equal size bins on the essays corpus, using self-reports. Models are the majority class baseline (Base); J48 decision tree (J48); Nearest neighbour (NN); Naive Bayes (NB); JRip rule set (JRIP); AdaboostM1 (ADA); Support vector machines (SMO).

Feature set comparison: In order to evaluate how each feature set contributes to the final result, we trained binary classifiers using the algorithms producing the best overall results with each feature set. We only analyse LIWC and MRC features for the essays corpus, as utterance type and prosodic features don't apply to written texts. We use the Naive Bayes, AdaboostM1 and SMO classifiers as they give the best performances with the full feature set. Results are shown in Table 13.

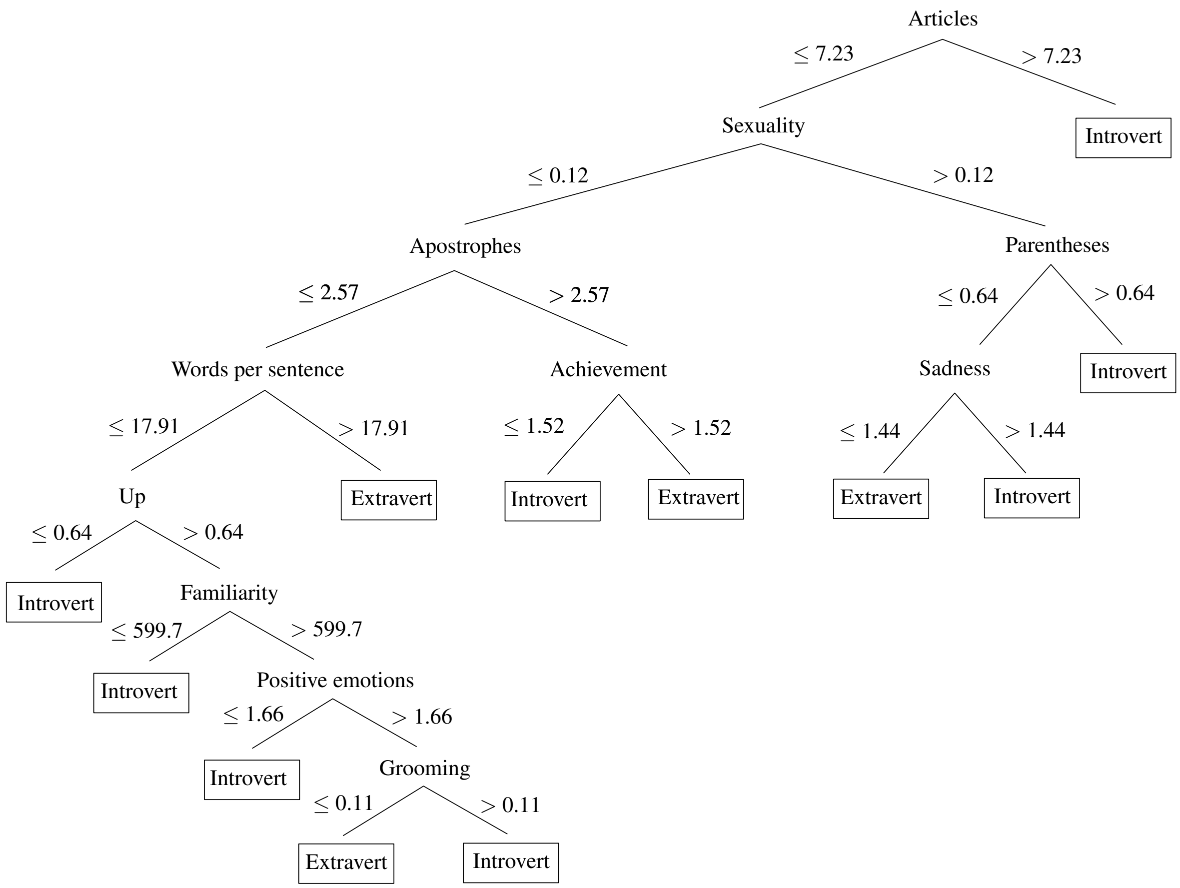

Figure 1: J48 decision tree for binary classification of extraversion, based on the essays corpus and self-reports.

Remarkably, we can see that the LIWC features outperform the MRC features for every trait, and the LIWC features on their own always perform slightly better than the full feature set. This clearly suggests that MRC features aren't as helpful as the LIWC features for classifying personality from written text, however Table 13 shows that they can still outperform the baseline for four traits out of five.

Concerning the algorithms, we find that AdaboostM1 performs the best for extraversion (56.3% correct classifications), while SMO produces the best models for all other traits. It suggests that support vector machines are promising for modelling personality in general. The easiest trait to model is still openness to experience, with 62.5% accuracy using LIWC features only.

4.2 EAR Corpus

Classification accuracies for the EAR corpus are in Table 14. We find that extraversion is the easiest trait to model using observer reports, with both Naive Bayes and AdaboostM1

Feature set	**None**	**LIWC features**			**MRC features**		
Classifier	Base	NB	ADA	SMO	NB	ADA	SMO
Set size	0	88	88	88	14	14	14
Extraversion	50.04	52.71	**56.34**•	52.75	52.87•	51.45	53.88
Emotional stability	50.08	56.02•	55.33•	**58.20**•	52.39	52.06	53.52•
Agreeableness	50.36	54.12•	52.71	**56.39**•	53.03•	52.06	53.31•
Conscientiousness	50.57	53.92•	54.48•	**55.62**•	53.03	52.95	53.84
Openness to experience	50.32	58.92•	58.64•	**62.52**•	55.41•	56.70•	57.47•

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 13: Classification accuracies with two equal size bins on the essays corpus using the majority class baseline (Base), Naive Bayes (NB), AdaboostM1 (ADA) and Support Vector Machine (SMO) classifiers, for different feature sets. Best model for each trait are in bold.

outperforming the baseline with an accuracy of 73.0%. The J48 decision tree for extraversion with a 66.8% accuracy is shown in Figure 2. Emotional stability is modelled with comparable success using a Naive Bayes classifier, however the improvement over the baseline is lower than with extraversion (22.8% vs. 25.2%) and other classifiers don't perform as well. Models of observed conscientiousness also outperform the baseline, with 67.7% accuracy using a Naive Bayes classifier, while the best model for agreeableness produces 61.3% correct classifications. None of the models for openness to experience significantly outperform the baseline, which suggests that openness to experience is expressed more clearly in stream of consciousness essays and self-reports than in the EAR dataset. Support vector machines don't perform as well as with the essays corpus, probably because of the sparseness of the dataset. Self-reports are much harder to model than observer reports given the same dataset size, as none of the self-report classifiers significantly outperform the majority class baseline.

Feature set comparison: For the EAR corpus we investigated the importance of all 4 feature sets: utterance type, LIWC, MRC, and prosodic features. We use the Naive Bayes models with the observer ratings as they perform the best with all features. Interestingly, Table 15 shows that the good classification accuracies for extraversion come from a combination of LIWC, MRC and prosodic features, as they all outperform the baseline on their own, but don't do as well as the 73.0% accuracy with the full feature set. Moreover, extraversion is the only trait for which prosody seems to make a difference. LIWC features are the main indicators of emotional stability, although the model with all features still performs better. MRC features are the most important for classifying conscientiousness (66.8%), while prosodic features produce the best model of openness to experience with 64.6% accuracy, improving on the model with all features. Although utterance type features never outperform the baseline on their own, the lack of significance could be the result of the small

1. Although equal size bins were used, the baseline accuracies differ from 50% because of the random sampling of the cross-validation.

Data	Trait	Base	J48	NN	NB	JRIP	ADA	SMO
Obs	Extra	47.78	66.78	59.33	73.00•	60.44	73.00 •	65.78
Obs	Emot	51.11	62.56	58.22	73.89•	56.22	48.78	60.33
Obs	Agree	47.78	48.78	51.89	61.33•	51.89	52.89	56.33
Obs	Consc	47.78	57.67	61.56	67.67•	61.56	60.22 •	57.11
Obs	Open	47.78	52.22	46.78	57.00	49.67	50.56	55.89
Self	Extra	47.78	48.78	49.67	57.33	50.56	54.44	49.89
Self	Emot	51.11	45.56	46.78	50.44	46.78	41.89	44.33
Self	Agree	52.22	47.89	50.89	58.33	56.89	55.22	52.33
Self	Consc	51.11	33.44	45.56	39.33	43.11	46.11	53.22
Self	Open	51.11	52.00	42.22	61.44	45.00	56.00	47.78

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 14: Classification accuracy with two equal size bins on the EAR corpus, for observer ratings (Obs) and self-reports (Self). Models are majority class baseline (Base)[1]; J48 decision tree (J48); Nearest neighbour (NN); Naive Bayes (NB); JRip rules set (JRIP); AdaboostM1 (ADA); Support vector machines (SMO).

Feature set	None	Type	LIWC	MRC	Prosody
Set size	0	4	88	14	11
Extraversion	47.78	45.67	68.89•	68.78•	67.56•
Emotional stability	51.11	60.22	69.89•	60.78	61.78
Agreeableness	47.78	57.56	54.00	58.67	50.44
Conscientiousness	47.78	59.67	60.22	66.78•	52.11
Openness to experience	47.78	53.11	61.11	54.00	**64.56•**

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 15: Classification accuracies for the EAR corpus with observer reports using the Naive Bayes classifier, for different feature sets (None=baseline, Type=utterance type). Models performing better than with the full feature set are in bold.

dataset size, since Section 3.3 showed that some utterance type features strongly correlate with several personality traits.

4.3 Qualitative Analysis

Decision trees and rule-based models can be easily understood, and can therefore help to uncover new linguistic markers of personality. Our models replicate previous findings, such as the link between verbosity and extraversion (c.f. *Word count* node of Figure 2), but they also provide many new markers.

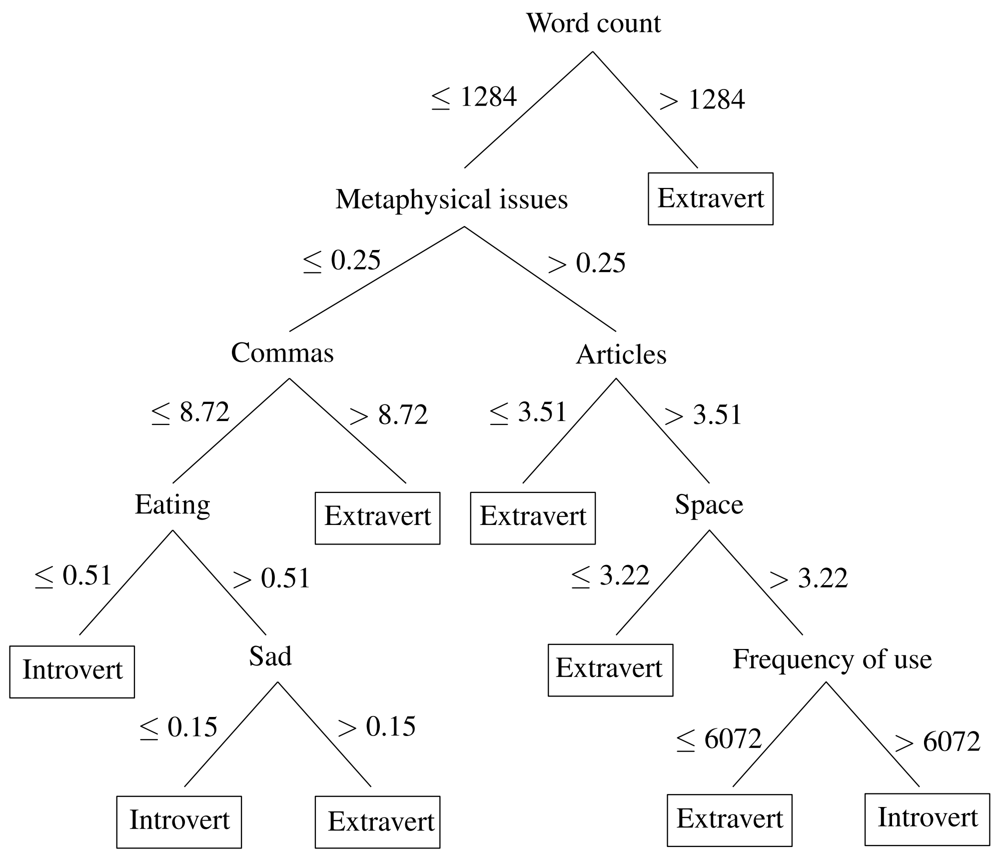

Figure 2: J48 decision tree for binary classification of extraversion, based on the EAR corpus and observer reports.

#	**Ordered rules**
1	(School $\geq$ 1.47) and (Motion $\geq$ 1.71) $\Rightarrow$ Not open
2	(Occup $\geq$ 2.49) and (Sixltr $\leq$ 13.11) and (School $\geq$ 1.9) and (I $\geq$ 10.5) $\Rightarrow$ Not open
3	(Fam $\geq$ 600.335106) and (Friends $\geq$ 0.67) $\Rightarrow$ Not open
4	(Nlet $\leq$ 3.502543) and (Number $\geq$ 1.13) $\Rightarrow$ Not open
5	(School $\geq$ 0.98) and (You $\leq$ 0) and (AllPct $\leq$ 13.4) $\Rightarrow$ Not open
6	Any other feature values $\Rightarrow$ Open

Table 16: JRip rule set for binary classification of openness to experience, based on the essays corpus.

The model of self-assessed openness to experience detailed in Table 16 shows that students referring a lot to school work tend to have low scores for that trait (Rules 1, 2 and 5). As expected, the avoidance of longer words is also indicative of a lack of creativity/conventionality (Rules 4 and 5), as well as the use of high-familiarity words and references to friends (Rule 3).

The model of observed extraversion in Figure 2 shows that word count is the most important feature for classifying that trait as an observer. The model also suggests that given low verbosity, extraversion can still manifest itself through the use of words related to metaphysical issues together with few articles, as well as through the use of many commas. The association between extraversion and the avoidance of articles probably reflects the use of more pronouns over common nouns and confirms previous findings associating extraversion with implicit language (Heylighen & Dewaele, 2002).

Interestingly, the decision tree trained on the essays corpus in Figure 1 for self-reported extraversion differs a lot from the observer model in Figure 2. While word count is the most important feature for observers, it doesn't seem to be a marker of self-assessed extraversion (see Section 3.3), although the number of words per sentence is used to discriminate on a subset of the data. On the other hand, the self-report model associates introversion with the use of articles, which was also the case in the observer model. While sexual content doesn't affect the observer model, it is the second most important feature for modelling self-reported extraversion. For example, participants using many sex-related words are modelled as introvert, unless they avoid parentheses and words related to sadness.

5. Regression Results

We also trained regression models using the same corpora. The baseline is a model returning the mean of all personality scores in the training set. We use the relative absolute error for evaluation, which is the ratio between the model's prediction error and the error produced by the baseline. A low relative error therefore indicates that the model performs better than the constant mean baseline, while a 100% relative error implies a performance equivalent to that baseline. All results are averaged over a 10-fold cross-validation, and all significance tests were done using a two-tailed paired t-test at the $p < .05$ level.

5.1 Essays Corpus

Regression results with the essays corpus and self-reports are in Table 17. Paired t-tests show that emotional stability and openness to experience produce models that significantly improve over the baseline. As with the classification task, openness to experience is the easiest trait to model using essays: four regression models out of five outperform the baseline. The M5' model tree produces the best result with a 93.3% relative error for openness to experience (6.7% error decrease), and a 96.4% relative error for emotional stability.

In terms of correlation between the model predictions and the actual ratings, the model for emotional stability and openness to experience produce Pearson's correlation coefficients of 0.24 and 0.33, respectively. Although the magnitude of the improvement seems relatively small, one needs to keep in mind the difficulty of the regression task over the binary classification task: it is the most fine-grained personality recognition problem, requiring the association of an exact scalar value with each individual.

Feature set comparison: Table 18 provides results for a comparison of LIWC with the MRC feature sets using the linear regression model, the M5' model tree and the support

Trait	Base	LR	M5R	M5	REP	SMO
Extraversion	100.00	99.17	99.31	99.22	99.98	100.65
Emotional stability	100.00	96.87•	99.75	96.43•	99.35	98.35
Agreeableness	100.00	98.92	99.86	99.22	99.78	100.28
Conscientiousness	100.00	98.68	100.62	98.56	100.47	99.30
Openness to experience	100.00	93.58•	97.68•	93.27•	99.82	94.19•

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 17: Relative error for regression models trained on the essays corpus with all features. Models are the mean value baseline (Base), linear regression (LR); M5' regression tree (M5R), M5' model tree with linear models (M5), REPTree (REP) and Support vector machines for regression (SMO).

vector machine algorithm for regression (SMOreg). Overall, LIWC features perform better than MRC features except for extraversion, for which the linear regression model with MRC features produces better results than with the full feature set. For all other traits, LIWC features on their own perform better than the full feature set, and almost always significantly outperform the baseline. The model for openness to experience produces the lowest relative error, with 6.50% improvement over the baseline.

Feature set	**None**	**LIWC features**			**MRC features**		
Regression model	Base	LR	M5	SMO	LR	M5	SMO
Extraversion	100.00	99.39	99.25•	100.8	**98.79•**	**98.79•**	99.13•
Emotional stability	100.00	96.71•	**96.42•**	98.03	99.49	99.54	99.89
Agreeableness	100.00	**98.50•**	98.52•	99.52	99.75	99.81	99.31•
Conscientiousness	100.00	98.23•	**98.14•**	99.46	99.23	99.23	99.16•
Openness to experience	100.00	**93.50•**	93.70•	94.14•	97.44•	97.44•	97.26•

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 18: Relative error for regression models trained on the essays corpus with the MRC and LIWC feature sets only. Models are linear regression (LR); M5' model tree (M5); Support vector machines for regression (SMO). Best models are in bold.

5.2 EAR Corpus

Regression results for the EAR corpus are in Table 19. A paired t-test (two-tailed, $p < .05$) over the cross-validation folds shows that the error reduction is significant for observed extraversion (79.9% relative error, i.e. 20.1% error decrease), conscientiousness (14.3% improvement) and emotional stability (13.3% improvement). While extraversion is the easiest trait to model from observer ratings, models of agreeableness and openness to experience don't outperform the baseline.

In terms of correlation between the model predictions and the actual ratings, the models for extraversion, emotional stability and conscientiousness respectively produce Pearson's correlation coefficients of 0.54, 0.47 and 0.44, significantly outperforming the baseline. Such correlations are relatively high, given that the average correlations between the ratings of each pair of observers is 0.54 for extraversion, 0.29 for emotional stability and 0.51 for conscientiousness (18 observers, between 31 and 33 data points for each pair).

Linear regression and support vector machines perform poorly, suggesting that they require a bigger dataset as in the essays corpus. As in the classification task, self-reports of the EAR corpus are clearly difficult to model: none of the models show significant improvement over the baseline.

Data	Trait	Base	LR	M5R	M5	REP	SMO
Obs	Extraversion	100.00	179.16	82.16 •	80.15	79.94 •	140.05
Obs	Emotional stability	100.00	302.74	92.03 •	86.75 •	100.51	162.05
Obs	Agreeableness	100.00	242.68	96.73	111.16	99.37	173.97
Obs	Conscientiousness	100.00	188.18	82.68 •	90.85	98.08	131.75
Obs	Openness to experience	100.00	333.65	101.64	119.53	102.76	213.20
Self	Extraversion	100.00	204.96	104.50	118.44	99.94	176.51
Self	Emotional stability	100.00	321.97	104.10	108.39	99.91	233.19
Self	Agreeableness	100.00	349.87	106.90	110.84	101.64	201.80
Self	Conscientiousness	100.00	177.12	103.39	120.29	107.33	124.91
Self	Openness to experience	100.00	413.70	107.12	122.68	126.31	233.01

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 19: Relative error for regression models, with observer ratings (Obs) and self-reports (Self) of the EAR corpus. Models are the mean value baseline (Base); linear regression (LR); M5' regression tree (M5R); M5' model tree with linear models (M5); REPTree decision tree (REPT); Support vector machines for regression (SMO). The relative error of the baseline model is 100%.

Feature set comparison: We trained regression models with each individual feature set using only observer reports, since self-reports didn't produce any significant result using all features. We only focus on the three regression tree algorithms as they perform the best with all features. Table 20 shows that LIWC are good predictors of observed extraversion, as the REPTree outperforms the same model with all features with a 76.4% relative error (23.6% improvement over the baseline). LIWC features also produce the best regression model for conscientiousness (82.1% relative error, 17.9% improvement). Surprisingly, the best model of emotional stability contains only prosodic features, with a 85.3% relative error (14.7% improvement). This finding suggests that speech cues are crucial for the perception of neuroticism, which could explain why Gill and Oberlander (2003) reported a low correlation between self-assessed and observed emotional stability using text only. As in the classification task, utterance type features don't show any significant improvement on their own.

Set	**Utterance type**			**LIWC features**			**MRC features**			**Prosodic features**		
Model	M5R	M5	REP	M5R	M5	REP	M5R	M5	REP	M5R	M5	REP
Extra	100.0	103.7	101.8	81.61	77.84•	**76.38•**	99.23	102.2	99.69	94.07	90.91	88.31•
Emot	102.5	103.0	102.6	90.79•	109.6	109.6	93.13•	96.08	104.4	92.24•	**85.32•**	97.95
Agree	102.4	102.7	111.1	98.49	111.7	102.5	104.1	112.5	102.2	100.0	108.4	108.9
Consc	100.0	95.04	104.1	**82.13•**	96.62	93.50	97.00	102.0	91.24•	100.0	104.7	101.7
Open	101.1	99.03	109.9	105.1	129.5	103.7	106.2	111.6	105.5	100.1	113.5	99.93

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 20: Relative error for regression models trained on the EAR corpus with individual feature sets. Models are M5' regression tree (M5R); M5' model tree with linear models (M5); REPTree regression tree (REP). Best models are in bold.

5.3 Qualitative Analysis

Regression trees for extraversion and conscientiousness are in Figures 3 and 4. As suggested by the correlations in Section 3.3, the model in Figure 3 shows that the voice's pitch and variation of intensity play an important role when modelling extraversion. A high verbal output is perceived as a sign of extraversion (see *Word Count* nodes), confirming previous findings (Scherer, 1979). On the other hand, a low mean pitch combined with a constant voice intensity characterises high introverts.

Figure 4 suggests that conscientious people use fewer swear words and content related to sexuality, while preferring longer words. The same figure also shows that conscientious people use fewer pronouns, i.e. a more explicit style, as well as more words related to communication (e.g., *talk* and *share*).

6. Ranking Results

Results with both corpora and different feature sets are in Tables 21 and 22. The models are trained over 100 rounds of boosting. The baseline model ranks extracts randomly, producing a ranking loss of 0.5 on average (lower is better). Results are averaged over a 10-fold cross-validation, and all significance tests were done using a two-tailed paired t-test at the $p < .05$ level.

6.1 Essays Corpus

Table 21 shows that openness to experience produces the best ranking model with the essays corpus, producing a ranking loss of 0.39 (lower is better). Remarkably, this trait was the easiest to model for all three recognition tasks with that corpus. As it is not the case with conversational data, it seems that streams of consciousness, or more generally personal writings, are likely to exhibit cues relative to the author's openness to experience. Emotional stability produces the second best model with a ranking loss of 0.42, followed by conscientiousness and extraversion, while the model for agreeableness produces the highest

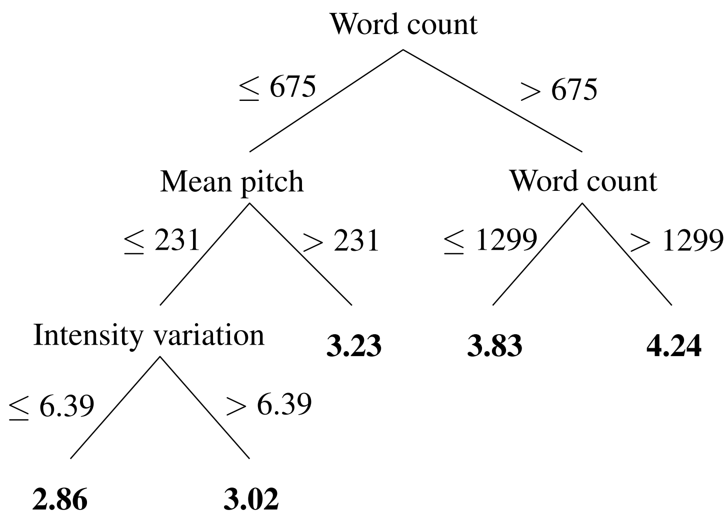

Figure 3: M5' regression tree for observed extraversion, computed using the EAR corpus. The target output ranges from 1 to 5.5, where 5.5 means strongly extravert (the highest value in the means of the observer ratings). The mean pitch value is expressed in Hertz, and the intensity variation (standard deviation) in decibels.

ranking loss. All models significantly outperform the random ranking baseline, but the actual improvement is still relatively small.

Feature set	**Base**	**All**	**LIWC**	**MRC**
Extraversion	0.50	0.44•	**0.44•**	0.46•
Emotional stability	0.50	0.42•	**0.42•**	0.47•
Agreeableness	0.50	0.46•	**0.46•**	0.48•
Conscientiousness	0.50	0.44•	**0.44•**	0.47•
Openness to experience	0.50	0.39•	**0.39•**	0.44•

• statistically significant improvement over the random ordering baseline (two-tailed paired t-test, $p < .05$)

Table 21: Ranking loss for the essays corpus over a 10-fold cross-validation for different feature sets and the random ordering baseline (Base). Best models are in bold (lower is better).

Feature set comparison: To evaluate which features contribute to ranking accuracy, we trained a ranking model with each feature set. Table 21 clearly shows that the LIWC features are the only contributors to model accuracy, as the inclusion of MRC features doesn't reduce the ranking loss for any trait.

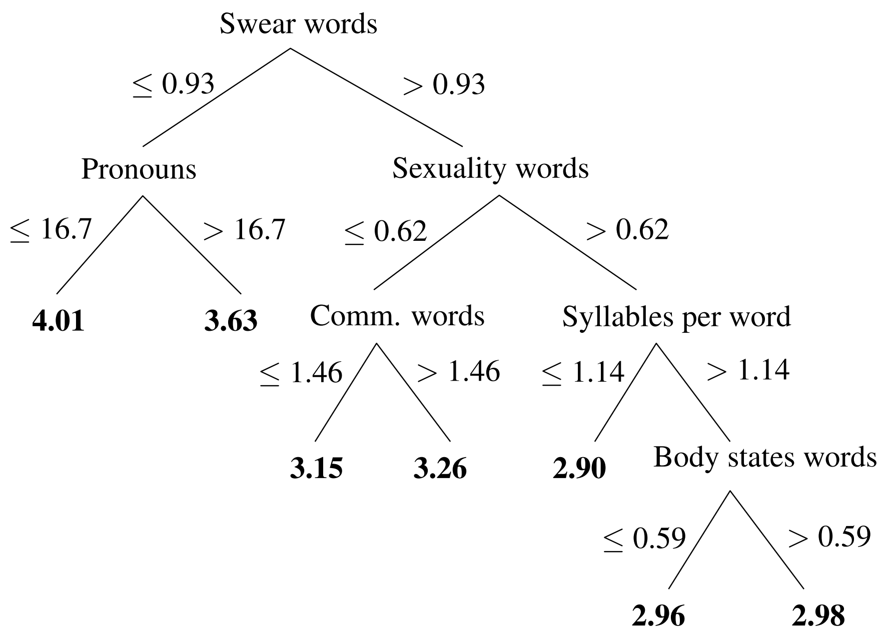

Figure 4: M5' regression tree for observed conscientiousness, computed using the EAR corpus. The target output ranges from 1 to 7, where 7 means strongly conscientious (*Comm. words* is the ratio of words related to communication).

6.2 EAR Corpus

Concerning the EAR corpus, Table 22 reporting experiments using all the features, shows that models of extraversion, agreeableness, conscientiousness, and openness to experience are better than the random ranking baseline. Emotional stability is the most difficult trait to model, while agreeableness and conscientiousness produce the best results, with ranking losses of 0.31 and 0.33 respectively.

Feature set	None	All	LIWC	MRC	Type	Prosody
Extraversion	0.50	0.35•	0.36•	0.45	0.55	**0.26•**
Emotional stability	0.50	0.41	0.41	**0.39•**	0.43	0.45
Agreeableness	0.50	**0.31•**	0.32•	0.44	0.45	0.54
Conscientiousness	0.50	**0.33•**	0.36•	0.41•	0.44	0.55
Openness to experience	0.50	0.38•	**0.37•**	0.41	0.49	0.44

• statistically significant improvement over the random ordering baseline (two-tailed paired t-test, $p < .05$)

Table 22: Ranking loss for the EAR corpus and observer reports[1] over a 10-fold cross-validation for different feature sets (None=baseline, Type=utterance type). Best models are in bold (lower is better).

Feature set comparison: When looking at individual feature sets, Table 22 shows that LIWC features perform significantly better than the baseline for all dimensions but emotional stability, while emotional stability is best predicted by MRC features only (0.39 ranking loss). Interestingly, prosodic features are very good predictors of extraversion, with a lower ranking error than the full feature set (0.26). This model produces the best overall result, with a 74% chance that the model will detect the most extravert among any two unseen conversation extracts. As in the previous recognition tasks, utterance type features on their own never significantly outperform the baseline.

6.3 Qualitative Analysis

The RankBoost rules indicate the impact of each feature on the recognition of a personality trait by the magnitude of the parameter α associated with that feature. Tables 23 to 25 show the rules with the most impact on each of the best models, with the associated α values. The feature labels are in Table 6. For example, the model of extraversion in Table 23 confirms previous findings by associating this trait with longer conversations (Rule 5), a high speech rate (Rules 1 and 4) and a high pitch (Rules 2, 6 and 7) (Nass & Lee, 2001). But new markers emerge, such as a high pitch variation for introverts (Rules 15, 18 and 20), contradicting previous findings reported by Scherer (1979).

Extraversion model with prosodic features					
#	**Positive rules**	α	#	**Negative rules**	α
1	Word-per-sec ≥ 0.73	1.43	11	Pitch-max ≥ 636.35	-0.05
2	Pitch-mean ≥ 194.61	0.41	12	Pitch-slope ≥ 312.67	-0.06
3	Voiced ≥ 647.35	0.41	13	Int-min ≥ 54.30	-0.06
4	Word-per-sec ≥ 2.22	0.36	14	Word-per-sec ≥ 1.69	-0.06
5	Voiced ≥ 442.95	0.31	15	Pitch-stddev ≥ 115.49	-0.06
6	Pitch-max ≥ 599.88	0.30	16	Pitch-max ≥ 637.27	-0.06
7	Pitch-mean ≥ 238.99	0.26	17	Pitch-slope ≥ 260.51	-0.12
8	Int-stddev ≥ 6.96	0.24	18	Pitch-stddev ≥ 118.10	-0.15
9	Int-max ≥ 85.87	0.24	19	Int-stddev ≥ 6.30	-0.18
10	Voiced ≥ 132.35	0.23	20	Pitch-stddev ≥ 119.73	-0.47

Table 23: Subset of the RankBoost model for extraversion with prosodic features only, based on EAR conversations and observer reports. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, i.e. introversion.

Concerning agreeableness, Rules 1 and 20 in Table 24 suggest that agreeable people use longer words but shorter sentences, and Rules 2 and 4 show that they express more tentativity (with words like *maybe* or *perhaps*) and positive emotions (e.g., *happy* and *good*). Anger and swear words greatly reduce the agreeableness score (Rules 12, 13, 18 and 19), as well as the use of negations (Rule 15).

1. We also built models of self-reports of personality based on the EAR corpus, but none of them significantly outperforms the baseline.

Agreeableness model with all features					
#	Positive rules	α	#	Negative rules	α
1	Nphon $\geq$ 2.66	0.56	11	Fam $\geq$ 601.61	-0.16
2	Tentat $\geq$ 2.83	0.50	12	Swear $\geq$ 0.41	-0.18
3	Colon $\geq$ 0.03	0.41	13	Anger $\geq$ 0.92	-0.19
4	Posemo $\geq$ 2.67	0.32	14	Time $\geq$ 3.71	-0.20
5	Voiced $\geq$ 584	0.32	15	Negate $\geq$ 3.52	-0.20
6	Relig $\geq$ 0.43	0.27	16	Fillers $\geq$ 0.54	-0.22
7	Insight $\geq$ 2.09	0.25	17	Time $\geq$ 3.69	-0.23
8	Prompt $\geq$ 0.06	0.25	18	Swear $\geq$ 0.61	-0.27
9	Comma $\geq$ 4.60	0.23	19	Swear $\geq$ 0.45	-0.27
10	Money $\geq$ 0.38	0.20	20	WPS $\geq$ 6.13	-0.45

Table 24: Best RankBoost model based on EAR conversations for agreeableness. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, i.e. disagreeableness.

Table 25 shows that conscientious people talk a lot about their work (Rule 1), while unconscientious people swear a lot (Rules 19). Insight words (e.g., *think* and *know*) are also good indicator of conscientiousness, as well as words expressing positive feelings like *happy* and *love* (Rule 2 and 3). Interestingly, conscientious people are modelled as having a high variation of their voice intensity (Rule 4). On the other hand, Rule 20 shows that speaking very loud produces the opposite effect, as well as having a high pitch (Rule 13). Long utterances are also indicative of a low conscientiousness (Rule 12).

The rule sets presented here contain only the most extreme rules of our ranking models, which contain many additional personality cues that aren't identified through a typical correlational analysis. For example, a high speech rate and a high mean pitch tend to contribute to a high extraversion ranking in Table 23's model, but they don't correlate significantly with observer ratings, as detailed in Table 11. Similarly, positive emotion words (Posemo) and the avoidance of long utterances (WPS) indicate agreeableness in the model in Table 24, while these features don't correlate significantly with agreeableness ratings.

7. Related Work

To our knowledge, there are only two other studies on the automatic recognition of personality. Both of these studies have focused on the classification of written texts based on self-reports, rather than using continuous modelling techniques as we do here.

Argamon et al. (2005) use the essays corpus of Pennebaker and King (1999), so their results are directly comparable to ours. As in our work, they use a top-down approach to feature definition: their feature set consists of relative frequencies of 675 function words and word categories based on networks of the theory of Systemic Functional Grammar. However, they simplify the task by removing the middle third of the dataset, thereby potentially increasing precision at the cost of reducing recall to a maximum of 67%. They train SMO models on the top third and lower third of the essays corpus for the two personality traits

Conscientiousness model with all features					
#	**Positive rules**	α	#	**Negative rules**	α
1	Occup $\geq$ 1.21	0.37	11	Swear $\geq$ 0.20	-0.18
2	Insight $\geq$ 2.15	0.36	12	WPS $\geq$ 6.25	-0.19
3	Posfeel $\geq$ 0.30	0.30	13	Pitch-mean $\geq$ 229	-0.20
4	Int-stddev $\geq$ 7.83	0.29	14	Othref $\geq$ 7.64	-0.20
5	Nlet $\geq$ 3.29	0.27	15	Humans $\geq$ 0.83	-0.21
6	Comm $\geq$ 1.20	0.26	16	Swear $\geq$ 0.93	-0.21
7	Nphon $\geq$ 2.66	0.25	17	Swear $\geq$ 0.17	-0.24
8	Nphon $\geq$ 2.67	0.22	18	Relig $\geq$ 0.32	-0.27
9	Nphon $\geq$ 2.76	0.20	19	Swear $\geq$ 0.65	-0.31
10	K-F-nsamp $\geq$ 329	0.19	20	Int-max $\geq$ 86.84	-0.50

Table 25: Best RankBoost model based on EAR conversations for conscientiousness. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, i.e. unconscientiousness.

of extraversion and emotional stability, achieving accuracies on this subset of the data of 58% for both traits.

We believe it is likely that personality recognition models need to be based on the full range of values to be useful in any practical application. Nevertheless, in order to do a direct comparison, we also removed the middle third of the essays dataset and trained an SMO classifier with the LIWC features. We obtain 57% classification accuracy for extraversion and 60% for emotional stability, whereas when the same algorithm is applied to the whole corpus, we obtain accuracies of 55% for extraversion and 57% for emotional stability, significantly outperforming the baseline (see Table 12). Using the EAR conversational data and observer reports, accuracies of our SMO models remain at 65% for extraversion but increase to 63% for emotional stability (see Table 14).

These results suggest that our feature set in combination with that of Argamon et al. could possibly improve performance, as both feature sets perform comparably. Using their features, Argamon et al. identify that relative frequencies of a set of function words are the best predictor for extraversion, suggesting that those that refer to norms and certainty are the most salient. Concerning emotional stability, the feature set characterising appraisal produces by far the best results. Appraisal features are relative frequencies of positive and negative words as well as frequencies of each category in the Attitude network (e.g., affect, appreciation, judgement, etc.). They find that neurotics tend to use more words related to negative appraisal and affect, but fewer appreciation appraisal words, suggesting that they focus more on their personal feelings.

Oberlander and Nowson (2006) follow a bottom-up feature discovery method by training Naive Bayes and SMO models for four of the Big Five traits on a corpus of personal weblogs, using n-gram features extracted from the dataset. In order to be able to compare with Argamon et al., they report experiments where they remove texts with non-extreme personality scores from their corpus, but they also report experiments applying classification algorithms to seven different ways of partitioning the whole corpus into classes, motivated as approximating a continuous modelling approach. Although, their results aren't directly comparable to ours because they are based on different corpora, we report the results that use all instances of the dataset, as we believe that discarding some of the test data increases precision at the cost of making recall unacceptably low.

When building Naive Bayes models using the most frequent bi-grams and tri-grams computed over the full corpus, Oberlander and Nowson (2006) find that the model of agreeableness is the only one outperforming the baseline (54% accuracy, no level of significance mentioned). When keeping only n-grams that are distinctive of two extreme sets of a given trait, accuracies range from 65% for extraversion to 72% for emotional stability. Finally, when applying an automatic feature selection algorithm to the filtered set, accuracies increase to range from 83% for emotional stability to 93% for agreeableness. When testing whether these models generalise to a different corpus of weblogs, Nowson and Oberlander (2007) report binary classification accuracies ranging from 55% for extraversion to 65% for conscientiousness. Interestingly, models trained on the most extreme instances of the original corpus seem to outperform models trained on the full corpus, although no level of significance is mentioned. These studies show that careful feature selection greatly improves classification accuracy, and that n-grams can be appropriate to model self-reports of personality, although, as Oberlander and Nowson point out, such features are likely to overfit. It would therefore be interesting to test in future work whether the feature sets used here generalise to another dataset.

Oberlander and Nowson (2006) also report results for 3-way and 5-way classification, in order to approximate the finer-grained continuous personality ratings used in psychology (as we do with the scalar models we present here). They obtain a maximum of 44.7% for extraversion with 5 bins, using raw n-grams (baseline is 33.8%). These results are not directly comparable to ours because they are on a different corpus, with different feature sets. Moreover, we have not provided results on such multiple classification experiments, because such models cannot take into account the fact that the different classes are part of a total ordering, and thus the resulting models are forced to ignore the importance of features that correlate with that ordering across all classes. We believe that regression and ranking models are more appropriate for finer-grained personality recognition (see Sections 5 and 6).

To evaluate this claim, we first mapped the output of the best classifier to a ranking and compared it with the RankBoost models. We trained a Naive Bayes classifier on the EAR corpus with observer reports and all features, using 5 equal size bins.[2] For each test fold of a 10-fold cross-validation, we computed the ranking loss produced by the classifier based on the ordering of the five classes. Results in Table 26 show that RankBoost significantly outperforms the classifier for four traits out of five ($p < .05$), with an improvement close to significance for emotional stability ($p = 0.12$).

Because RankBoost's goal is to minimise the ranking loss, this comparison is likely to favour ranking models. Therefore, we also mapped the output of the RankBoost models to 5 classification bins to see whether RankBoost could perform as well as a classifier for the classification task. We divided the output ranking into 5 bins, each containing a 20% slice of contiguously ranked instances. Results in Table 26 show that the Naive Bayes classifier never outperforms RankBoost significantly, while the ranking model produces a

2. Oberlander and Nowson use unequal bins defined for each personality trait using standard deviation from the mean, which may be an easier task than equal size bins.

Task	Ranking			Classification		
Model	Base	NB	Rank	Base	NB	Rank
Extraversion	0.50	0.48	0.35•	20.0	32.3	32.1
Emotional stability	0.50	0.50	0.41	20.0	21.9	21.9
Agreeableness	0.50	0.50	0.31•	20.0	28.4	37.8
Conscientiousness	0.50	0.46	0.33•	20.0	34.7	30.3
Openness to experience	0.50	0.53	0.38•	20.0	19.8	26.8

• statistically significant improvement over the other model (two-tailed t-test, $p < .05$)

Table 26: Comparison between ranking (Rank) and classification models (NB) for both personality ranking and classification tasks (5 bins). Evaluation metrics are ranking loss (lower is better) and classification accuracy (higher is better), respectively. Results are averaged over a 10-fold cross-validation.

better mean accuracy for agreeableness (38%) and openness to experience (27%), and the same accuracy for emotional stability (22%). In sum, we find that ranking models perform as well for classification and better for ranking compared with our best classifier, thus modelling personality using continuous models is more accurate.

8. Discussion and Future Work

We show that personality can be recognised by computers through language cues.[3] While recent work in AI explores methods for the automatic detection of other types of pragmatic variation in text and conversation, such as opinion, emotion, and deception, to date, we know of only two studies besides our own on *automatic recognition* of user personality (Argamon et al., 2005; Mairesse & Walker, 2006a, 2006b; Oberlander & Nowson, 2006). To our knowledge, the results presented here are the first to demonstrate statistically significant results for texts and to recognise personality in conversation (Mairesse & Walker, 2006a, 2006b). We present the first results applying regression and ranking models in order to model personality recognition using the continuous scales traditional in psychology. We also systematically examine the use of different feature sets, suggested by previous psycholinguistic research. Although these features have been suggested by the psycholinguistic literature, reported correlations with personality ratings are generally weak: it was not obvious that they would improve accuracies of statistical models on unseen subjects.

Computational work on modelling personality has primarily focused on methods for *expressing* personality in virtual agents and tutorial systems, and concepts related to personality such as politeness, emotion, or social intelligence (Walker, Cahn, & Whittaker, 1997; André, Klesen, Gebhard, Allen, & Rist, 1999; Lester, Towns, & FitzGerald, 1999; Wang, Johnson, Mayer, Rizzo, Shaw, & Collins, 2005) *inter alia*. Studies have shown that user evaluations of agent personality depend on the user's own personality (Reeves & Nass, 1996; Cassell & Bickmore, 2003), suggesting that an ability to model the user's personality

3. An online demo and a personality recognition tool based on the models presented in this paper can be downloaded from `www.dcs.shef.ac.uk/cogsys/recognition.html`

is required. Models such as we present here for the automatic recognition of user personality is one way to acquire such a user model (Chu-Carroll & Carberry, 1994; Thompson, Göker, & Langley, 2004; Zukerman & Litman, 2001). We plan to test these models as user models in the context of an adaptive dialogue system.

Table 27 summarises results for all the personality traits and recognition tasks we analysed. What clearly emerges is that extraversion is the easiest trait to model from spoken language, followed by emotional stability and conscientiousness. Concerning written language, models of openness to experience produce the best results for all recognition tasks. We can also see that feature selection is very important, as some of the best models only contain a small subset of the full feature set. Prosodic features are important for modelling observed extraversion, emotional stability and openness to experience. MRC features are useful for models of emotional stability, while LIWC features are beneficial for all traits. We also analysed qualitatively which features had the most influence in specific models, for all recognition tasks, as well as reporting correlations between each feature and personality traits in Section 3.3.

Although the parameters of the algorithms have not been optimised, the bottom of Table 27 seems to indicate that simple models like Naive Bayes or regression trees tend to outperform more complex ones (e.g., support vector machines), confirming results from Oberlander and Nowson (2006). However, our experiments on the larger essays corpus (more than 2,400 texts) show that support vector machines and boosting algorithms produce higher classification accuracies. It is therefore likely that those algorithms would also perform better on spoken data if they were trained on a much larger corpus than the EAR dataset, and if their parameters were optimised.

We hypothesised that models of observed personality will outperform models of self-assessed personality. Our results do suggest that observed personality may be easier to model than self-reports, at least in conversational data. For the EAR corpus, we find many good results with models of observed personality, while models of self-assessed personality never outperform the baseline. This may be due to objective observers using similar cues as our models, while self-reports of personality may be more influenced by factors such as the desirability of the trait (Edwards, 1953). Hogan (1982) introduced the distinction between the agent's and the observer's perspective in personality assessment. While the agent's perspective conceptually taps into a person's identity (or 'personality from the inside'), the observer's perspective in contrast taps into a person's reputation (or 'personality from the outside'). Both facets of personality have important psychological implications. A person's identity shapes the way the person experiences the world. A person's reputation, however, is psychologically not less important: it determines whether people get hired or fired (e.g., reputation of honesty), get married or divorced, get adored or stigmatised. Because it is harder to assess, this observer's perspective has received comparatively little attention in psychology. Given that in everyday life people act as observers of other people's behaviours most of the time, the external perspective naturally has both high theoretical importance and social relevance (Hogan, 1982).

Recent research exploring this issue in psychology is based on the Brunswikian Lens model (Brunswik, 1956), which has been used extensively in recent years to explain the 'kernel of truth' in the social perception of strangers. Use of the lens model in personality research reflects the widely shared assumptions that the expression of personality is commu-

Task	Classification			Regression			Ranking		
Baseline	n/a	none	50%	n/a	none	0%	n/a	none	0.50
Self-report models trained on written data (essays):									
Extraversion	ADA	LIWC	56%	LR	MRC	1%	Rank	LIWC	0.44
Emotional stability	SMO	LIWC	58%	M5	LIWC	4%	Rank	LIWC	0.42
Agreeableness	SMO	LIWC	56%	LR	LIWC	2%	Rank	LIWC	0.46
Conscientiousness	SMO	LIWC	56%	M5	LIWC	2%	Rank	LIWC	0.44
Openness to experience	SMO	LIWC	63%	M5	all	7%	Rank	LIWC	0.39
Observer report models trained on spoken data (EAR):									
Extraversion	NB	all	73%	REP	LIWC	24%	Rank	prosody	0.26
Emotional stability	NB	all	74%	M5	prosody	15%	Rank	MRC	0.39
Agreeableness	NB	all	61%	M5R	all*	3%	Rank	all	0.31
Conscientiousness	NB	all	68%	M5R	LIWC	18%	Rank	all	0.33
Openness to experience	NB	prosody	65%	M5	type*	1%	Rank	LIWC	0.37

Table 27: Comparison of the best models for each trait, for all three recognition tasks. Each table entry contains the algorithm, the feature set, and the model performance. See Sections 3.2 and 3.4 for details. Depending on the task, the evaluation metric is either the (1) classification accuracy; (2) percentage of improvement over the regression baseline; (3) ranking loss. Asterisks indicate results that aren't significant at the $p < .05$ level.

nicatively functional, i.e. that (a) latent attributes of persons are expressed via observable cues; (b) observers rely on observable cues to infer the latent attributes of others; (c) observers use appropriate cues – that is, their implicit assumptions on the relations between observable cues and latent attributes are to some extent accurate. The model has also been useful in identifying observable cues that mediate convergences between judgments of latent attributes and more direct measures of those attributes (Scherer, 2003; Heinrich & Borkenau, 1998).

As there are discrepancies between markers of self-assessed and observed personality, another issue is the identification of the most appropriate model given a specific application. Such a gold standard can be approximated by either observer or self-reports, however it is likely that for a specific trait one type of report will be closer to the *true* personality. A hypothesis that remains to be tested is that traits with a high visibility (e.g., extraversion) are more accurately assessed using observer reports, as they tend to yield a higher inter-judge agreement (Funder, 1995), while low visibility traits (e.g., emotional stability) are better assessed by oneself. A personality recogniser aiming to estimate the true personality would therefore have to switch from observer models to self-report models, depending on the trait under assessment.

Beyond practical applications of personality recognition models, this work is also an attempt to explore different ways of looking at the relation between personality and language. We looked at various personality recognition tasks, and applied different learning methods in Section 3.4. The tasks vary in complexity: a ranking model can be directly derived from a regression model, while a classification model can be derived from either a ranking or a regression model. Is any type of model closer to the actual relation between language, and more generally behaviour, and personality? Does personality vary continuously, or are there clusters of people with similar trait combinations? If the relation is continuous, classification algorithms will never be able to produce accurate models for more than two classes, because they don't take into account any ordering between the classes. As ranking models outperform classifiers (see Section 7), and given the wide range of individual differences reflected by the literature on the Big Five (Allport & Odbert, 1936; Norman, 1963; Goldberg, 1990), we believe that personality varies continuously among members of the population, suggesting that regression or ranking models should be more accurate in the long run. This hypothesis is supported by recent work in medical research showing that antisocial personality disorder varies continuously (Marcus, Lilienfeld, Edens, & Poythress, 2006). Regression provides the most detailed model of the output variables, but depending on whether absolute differences between personality scores are meaningful, or if only relative orderings between people matter, ranking may be more appropriate. Additional models could also be tried on the ranking task, such as support vector algorithms for ordinal regression (Herbrich, Graepel, & Obermayer, 2000). Moreover, future work should assess whether optimising the parameters of the learning algorithms improves performance.

In future work, we would like to improve these models and examine how well they perform across dialogue domains. It is not clear whether the accuracies are high enough to be useful. Applications involving speech recognition will introduce noise in all features except for the prosodic features, probably reducing model accuracy, but since the EAR corpus is relatively small, we expect that more training data would improve performance. Additionally, we believe that the inclusion of gender as a feature would produce better models, as the actual language correlates of perceived personality were shown to depend on the gender of the speaker (Mehl et al., 2006). We also believe that future work should investigate the combination of individual features in a trait-dependent way. Another issue is the poor performance of the utterance type features—since there were significant correlation results for these features in Section 3.3, it is unclear why these features are not useful in the statistical models. This could possibly arise from the small size of the datasets, or from the relatively low accuracy of our hand-crafted automatic tagger, compared to other work using supervised learning methods (Stolcke, Ries, Coccaro, Shriberg, Bates, Jurafsky, Taylor, Martin, Ess-Dykema, & Meteer, 2000; Webb, Hepple, & Wilks, 2005).

We have begun to test these models on our spoken language generator (Mairesse & Walker, 2007). In future work, we plan to compare the utility of models trained on out-of-domain corpora, such as those here, with other methods for training such models, in terms of their utility for the automatic adaptation of the output generation of dialogue systems.

Acknowledgments

We would like to thank James Pennebaker for giving us access to the essays data. This work was partially funded by a Royal Society Wolfson Research Merit Award to Marilyn Walker, and by a Vice Chancellor's studentship to François Mairesse.

References

Allport, G. W., & Odbert, H. S. (1936). Trait names: a psycho-lexical study. *Psychological Monographs*, *47*(1, Whole No. 211), 171–220.

André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (1999). Integrating models of personality and emotions into lifelike characters. In *Proceedings of the Workshop on Affect in Interactions - Towards a new Generation of Interfaces*, pp. 136–149.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. (2005). Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: a meta-analysis. *Journal of Applied Psychology*, *89*(5), 901–910.

Breck, E., Choi, Y., & Cardie, C. (2007). Identifying expressions of opinion in context. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA.

Byrne, D., & Nelson, D. (1965). Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology*, *1*, 659–663.

Cassell, J., & Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, *13*, 89–132.

Chu-Carroll, J., & Carberry, S. (1994). A plan-based model for response generation in collaborative task-oriented dialogue. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI)*, pp. 799–805.

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

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL.

Dewaele, J.-M., & Furnham, A. (1999). Extraversion: the unloved variable in applied linguistic research. *Language Learning*, *49*(3), 509–544.

Donnellan, M. B., Conger, R. D., & Bryant, C. M. (2004). The Big Five and enduring marriages. *Journal of Research in Personality*, *38*, 481–504.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that it will be endorsed. *Journal of Applied Psychology*, *37*, 90–93.

Enos, F., Benus, S., Cautin, R., Graciarena, M., Hirschberg, J., & Shriberg, E. (2006). Personality factors in human deception detection: Comparing human to machine performance. In *Proceedings of ICSLP*.

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3? criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*(8), 773–790.

Fast, L. A., & Funder, D. C. (2007). Personality as manifest in word use: Correlations with self-report, acquaintance-report, and behavior. *Journal of Personality and Social Psychology*, *in press*.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (1998). An efficient boosting algorithm for combining preferences. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 170–178.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670.

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, *64*(3), 479–490.

Furnham, A. (1990). Language and personality. In Giles, H., & Robinson, W. (Eds.), *Handbook of Language and Social Psychology*. Winley.

Furnham, A., Jackson, C. J., & Miller, T. (1999). Personality, learning style and work performance. *Personality and Individual Differences*, *27*, 1113–1122.

Furnham, A., & Mitchell, J. (1991). Personality, needs, social skills and academic achievement: A longitudinal study. *Personality and Individual Differences*, *12*, 1067–1073.

Gill, A., & Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 456–461.

Gill, A. (2003). *Personality and Language: The Projection and Perception of Personality in Computer-Mediated Communication*. Ph.D. thesis, University of Edinburgh.

Gill, A. J., & Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 363–368.

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

Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., & Kajarekar, S. (2006). Combining prosodic, lexical and cepstral systems for deceptive speech detection. In *Proceedings of IEEE ICASSP*.

Heinrich, C. U., & Borkenau, P. (1998). Deception and deception detection: The role of cross-modal inconsistency. *Journal of Personality*, *66*(5), 687–712.

Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In Smola, A. J., Bartlett, P., Schölkopf, B., & Schuurmans, D. (Eds.), *Advances in Large Margin Classifiers*, pp. 115–132. MIT Press, Cambridge, MA.

Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: an empirical measure. *Context in Context, Special issue of Foundations of Science*, *7*(3), 293–340.

Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B., Shriberg, E., & Stolcke, A. (2005). Distinguishing deceptive from non-deceptive speech. In *Proceedings of Interspeech'2005 - Eurospeech*.

Hogan, R. (1982). A socioanalytic theory of personality. *Nebraska Symposium of Motivation*, *30*, 55–89.

Hogan, R., Curphy, G. J., & Hogan, J. (1994). What we know about leadership: Effectiveness and personality. *American Psychologist*, *49*(6), 493–504.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The "Big Five" Inventory: Versions 4a and 5b. Tech. rep., Berkeley: University of California, Institute of Personality and Social Research.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A., & John, O. P. (Eds.), *Handbook of personality theory and research*. New York: Guilford Press.

Komarraju, M., & Karau, S. J. (2005). The relationship between the Big Five personality traits and academic motivation. *Personality and Individual Differences*, *39*, 557–567.

Lester, J. C., Towns, S. G., & FitzGerald, P. J. (1999). Achieving affective impact: Visual emotive communication in lifelike pedagogical agents. *The International Journal of Artificial Intelligence in Education*, *10*(3-4), 278–291.

Liscombe, J., Venditti, J., & Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Interspeech'2003 - Eurospeech*.

Mairesse, F., & Walker, M. (2006a). Automatic recognition of personality in conversation. In *Proceedings of HLT-NAACL*.

Mairesse, F., & Walker, M. (2006b). Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 543–548.

Mairesse, F., & Walker, M. (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 496–503.

Mallory, P., & Miller, V. (1958). A possible basis for the association of voice characteristics and personality traits. *Speech Monograph*, *25*, 255–260.

Marcus, D. K., Lilienfeld, S. O., Edens, J. F., & Poythress, N. G. (2006). Is antisocial personality disorder continuous or categorical? A taxometric analysis. *Psychological Medicine*, *36*(11), 1571–1582.

McLarney-Vesotski, A. R., Bernieri, F., & Rempala, D. (2006). Personality perception: A developmental study. *Journal of Research in Personality*, *40*(5), 652–674.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*, 862–877.

Mehl, M., Pennebaker, J., Crow, M., Dabbs, J., & Price, J. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, and Computers*, *33*, 517–523.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

Nass, C., & Lee, K. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171–181.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, *29*, 665–675.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, *66*, 574–583.

Nowson, S., & Oberlander, J. (2007). Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proceedings of the International Conference on Weblogs and Social Media*.

Nunn, S. (2005). Preventing the next terrorist attack: The theory and practice of homeland security information systems. *Journal of Homeland Security and Emergency Management*, *2*(3).

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, *42*, 239–270.

Oberlander, J., & Nowson, S. (2006). Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Oudeyer, P.-Y. (2002). Novel useful features and algorithms for the recognition of emotions in speech. In *Proceedings of the 1st International Conference on Speech Prosody*, pp. 547–550.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 115–124.

Paunonen, S. V., & Jackson, D. N. (2000). What is beyond the Big Five? plenty!. *Journal of Personality*, *68*(5), 821–836.

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptor. *Journal of Personality and Social Psychology*, *57*(3), 552–567.

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

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*, 1296–1312.

Pennebaker, J. W., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of HTL-EMNLP*.

Reeves, B., & Nass, C. (1996). *The Media Equation*. University of Chicago Press.

Reiter, E., & Sripada, S. G. (2004). Contextual influences on near-synonym choice. In *Proceedings of the International Natural Language Generation Conference*, pp. 161–170.

Rienks, R., & Heylen, D. (2006). Dominance detection in meetings using easily obtainable features. In Bourlard, H., & Renals, S. (Eds.), *Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Vol. 3869 of *Lecture Notes in Computer Science*. Springer Verlag.

Riggio, R. E., Salinas, C., & Tucker, J. (1988). Personality and deception ability. *Personality and Individual Differences*, *9*(1), 189–191.

Rosenberg, A., & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of Interspeech'2005 - Eurospeech*.

Rushton, J. P., Murray, H. G., & Erdle, S. (1987). Combining trait consistency and learning specificity approaches to personality, with illustrative data on faculty teaching performance. *Personality and Individual Differences*, *8*, 59–66.

Schapire, R. (1999). A brief introduction to boosting. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, *2*, 1401–1406.

Scherer, K. R. (1979). Personality markers in speech. In Scherer, K. R., & Giles, H. (Eds.), *Social markers in speech*, pp. 147–209. Cambridge University Press.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227–256.

Siegman, A., & Pope, B. (1965). Personality variables associated with productivity and verbal fluency in the initial interview. In *Proceedings of the 73rd Annual Conference of the American Psychological Association*.

Sigurdsson, J. F. (1991). Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. *Personality and Individual Differences*, *12*(6), 617–624.

Smith, B. L., Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Effects of speech rate on personality perception. *Language and Speech*, *18*, 145–152.

Somasundaran, S., Ruppenhofer, J., & Wiebe, J. (2007). Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–371.

Stoyanov, V., Cardie, C., & Wiebe, J. (2005). Multi-perspective question answering using the OpQA corpus. In *Proceedings of HLT-EMNLP*.

Thompson, C. A., Göker, M. H., & Langley, P. (2004). A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, *21*, 393–428.

Tucker, S., & Whittaker, S. (2004). Accessing multimodal meeting data: Systems, problems and possibilities. *Lecture Notes in Computer Science, Machine Learning for Multimodal Interaction*, *3361*, 1–11.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 417–424.

Vogel, K., & Vogel, S. (1986). L'interlangue et la personalité de l'apprenant. *International Journal of Applied Linguistics*, *24*(1), 48–68.

Walker, M., Cahn, J. E., & Whittaker, S. J. (1997). Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the 1st Conference on Autonomous Agents*, pp. 96–105.

Walker, M., & Whittaker, S. (1990). Mixed initiative in dialogue: an investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 70–78.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2005). The politeness effect: Pedagogical agents and learning gains. *Frontiers in Artificial Intelligence and Applications*, *125*, 686–693.

Watson, D., & Clark, L. A. (1992). On traits and temperament: General and specific factors of emotional experience and their relation to the five factor model. *Journal of Personality*, *60*(2), 441–76.

Webb, N., Hepple, M., & Wilks, Y. (2005). Error analysis of dialogue act classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistic*, *30*(3), 277–308.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, pp. 761–769.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA.

Zukerman, I., & Litman, D. (2001). Natural language processing and user modeling: Synergies and limitations. *User Modeling and User-Adapted Interaction*, *11*(1-2), 129–158.