Looking forward to more Extraversion with N-grams

Looking forward to more Extraversion with N-grams

Alastair J. Gill and Jon Oberlander

School of Informatics, University of Edinburgh

2 Buccleuch Place, Edinburgh, EH8 9LW UK

agill@cogsci.ed.ac.uk J.Oberlander@ed.ac.uk

Abstract. We study how Extraversion or Introversion influences people's language production. Extending recent work, we show how the use of larger-scale co-occurrences of words distinguishes these personality groups. Along with previous findings, our results suggest that Extraverts could be "lazy" and use collocations of words to economise on discourse planning. We compare these results with previous findings for personality language. Implications of using co-occurrence techniques are discussed.

1. Introduction

We study the impact of personality on textual communication, in particular through computer-mediated means. The trait Extraversion-Introversion is especially relevant since this describes sociability, which is important for communication, and is readily perceived, even in computer-mediated communication (Gill and Oberlander, 2003).

Recent work using the MRC psycholinguistic database has shown that Extraverts use words which are less *concrete*, and more abstract, like *thoughts, flavours, pains*, rather than referring to entities which can be sensed like *table, spoon, girl* (Gill and Oberlander, 2002). In addition, we went on to demonstrate that Extravert and Introvert authors are distinguished by a range of two-word collocations (bigrams). A summary of these features can be found in Figure 1.

Surface Realisation: Extraverts are more informal, use *hi*, and use looser punctuation (!! or ...); Introverts use *hello*.

Quantification: Introverts show greater use of quantifiers (for exaggeration?); Extraverts are looser and less specific.

Social Devices: Stylistic expressions such as *catch up* and *take care* indicate the Extravert's relaxed social style.

Self/Other: Reference Introverts use more first-person singular (*i*), whereas Extraverts are more likely to use plural *we*.

Valence: Introverts prominently use negations; Extraverts use words suggestive of positive affect.

Ability: Extraverts are more confident and assertive (eg., *want-*, *able-*, *need-*(*to*)); Introverts are more tentative and timid (*trying-*, *going-*(*to*)).

Modality: Extraverts are more strongly predictive than Introverts (eg., modal auxiliaries *will-* vs. *should-*(*be*)).

Message Planning/Expression: Introverts prefer co-ordinating conjunctions (*and*, *but*), whereas only Extraverts use the subordinative *which* (usually for evaluation?).

Figure 1: Extravert and Introvert Language

So far, these two separate findings have been viewed in isolation. However, in this paper we aim to draw them together in an explanation of Extravert discourse behaviour. We propose that Extraverts direct resources away from precise lexical planning, in an endeavour to construct utterances more quickly. Their drive to seize the conversational floor leads to a certain linguistic *laziness*. This is, however, not laziness in the sense of indolence. Rather, it is an *efficiency of action*, whereby new or precise linguistic decisions are avoided in favour of pre-existing, remembered choices. In particular, such speakers are more likely to rely upon stereotypical expressions and previously used or pre-planned chunks of language: The collocations found in the previous bigram analysis suggests that Extraverts use regularly co-occurring pairs of words more frequently than Introverts.

To test this theory, we build upon the bigram analysis, and extend it so as to consider larger collocations of words. The structure of the paper is as follows: First we will introduce in more detail the concept of Extraversion and why it is such an important personality trait. We then briefly describe some findings for Extravert language use. Next, we introduce the experimental method used for the original bigram analysis and detail the extensions used in the current analysis. Then follows the discussion and conclusion.

1.1 The importance of being Extravert

Intuitively, we get the impression that Extraverts tend to talk loudly and say more, whereas Introverts are more softly spoken and reserved. Are such hypotheses borne out by fact, and how else does this personality dimension influence language production? Before approaching this question, we define more precisely what is meant by Extraversion, and why this trait is important.

Extraversion is a trait which is strongly related to interpersonal interaction and sociability, and as a result there is a greater awareness of this trait and its manifestation in behaviour. A typical Extravert is described as someone who is sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert is quiet, retiring, reserved, plans ahead, and dislikes excitement (Eysenck and Eysenck, 1991).

The trait of Extraversion is central to the two major theories of personality psychology: Eysenck's three factor model; and the five factor model developed by Costa and McCrae and others (Matthews and Deary, 1998). Indeed, the personality trait of Extraversion is one of the few which researchers generally agree provides 'consistent and valid information' (Jonassen and Grabowski, 1993).

Despite the general agreement for the inclusion of Extraversion in personality theory, beyond this there is greater debate. For example, Eysenck's model of personality incorporates just two further dimensions: Neuroticism, which is mainly characterised by susceptibility to anxiety; and Psychoticism, which is more complicated, but generally related to aggression and individuality. By contrast, the NEO-PI-R model incorporates five factors (Costa and McCrae, 1992). In addition to Extraversion and Neuroticism, they proposed three other traits: Conscientiousness, Agreeableness and Openness, which are generally regarded as relating to Psychoticism; but this is still a matter of some debate (cf. Matthews and Deary, 1998).

But how does Extraversion influence an individual's language production? In addressing this question, we first outline some hypotheses from the literature, before describing our collection of a controlled corpus of language, and our analysis of it.

1.2 Previous hypotheses

From an intuitive perspective, Extraverts are described as individuals who think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic. Conversely, Introverts monopolise the conversation on topics important to them, are more self-focussed and prefer to concentrate on discussing one topic in depth (cf. Carment, Miles, and Cervin, 1965). With reference primarily to speech, Furnham (1990) has proposed that Extravert language is less formal, has a more restricted code, uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions), and uses vocabulary loosely (see also Dewaele and Furnham, 1999, for a review of speech and writing studies).

Text analysis approaches have found that transcribed texts rated as belonging to the *warm* facet of Extraversion used fewer negative emotion words and unique words, and more present tense verbs, with *dominant* texts using fewer unique words, positive emotion words and self referents (Berry, Pennebaker, Mueller, and Hiller, 1997). Finally, study of the texts *written* by Extraverts has found that they used fewer negations, tentative words, negative emotion words, causation words, inclusive words, and exclusive words, while using more social and positive emotion words (Pennebaker and King, 1999).

2. Method

Our extension of *n*-gram analysis uses the same data and methods as our previously reported bigram analysis (Gill and Oberlander, 2002), namely: 210 texts produced by 105 University students or recent graduates (37 males, 68 females; mean age = 24.3 years; SD = 4.6; all native English speakers) of known personality (EPQ Revised short form; Eysenck, Eysenck, and Barrett, 1985; mean score = 7.91, SD = 3.25; normative score = 7.42 (male), 7.60 (female)). Note that these personality scores depend on subjects' self-assessment: they do not depend on peer-judgement, and hence do not depend on external judgments concerning the subjects' verbal behaviours. Each participant composed two e-mails *to a good friend whom they hadn't seen for quite some time*, spending around 10 minutes on each message. The first e-mail concerned their activities in the past week, the second discussed their plans for the next week. The total corpus size is around 65,000 words.

The original corpus of texts was divided by degree of Extraversion by selecting respondents whose E score was greater or less than 1 s.d. of the mean (cf. Dewaele and Pavlenko, 2002), with the 21 High Extravert authors scoring more than 11, and the 17 Low Extravert authors scoring less than 5. The resulting Extravert and Introvert sub-corpora contain around 12,000 words and 8,000 respectively, which resulted from the average length of Extravert texts being longer than that of Introvert texts (around 570 words versus 470 words). These subcorpora were used for the subsequent calculation of *n*-grams. This was performed using word co-occurrence window lengths of 3 and 5 words.[1]

The trigram data for each corpus was then ranked by their co-occurrence significance using the log-likelihood statistic (-2 logλ), since for smaller corpora this approximates better to χ^2 than the X^2 statistic (Dunning, 1993). Rankings for each group are based on the top 50 trigrams with frequency of N >= 2, and a significance of p<.001. Relative frequency ratios (Damerau, 1993)[2] were then calculated for trigrams that were common to both the sub-

[1] Ted Pedersen's n-gram software is available from: http://www.d.umn.edu/~tpederse/code.html
[2] Note here that functors and rarer collocations are retained.

corpora, and a Spearman Rank correlation was then performed on this data. Note that here the *n*-gram analysis and relative frequency ratios are used for different purposes than those of, for example, Damerau (1993), who uses them to distinguish texts on the basis of key words. Due to a scarcity of data and statistical tools for 5-grams, frequency and relative frequency alone were calculated.

3. Results

3.1 Spearman Rank Correlation

Extravert and Introvert use of the shared trigrams is not significantly correlated r_S =.236 (N=13) at the $p < .05$ and therefore indicates that the two groups' usage of these is distinct.

3.2 N-grams

The results of the relative frequency ratio analysis of the trigrams, and those unique to Extravert and Introvert corpora can be found in Tables 1, 2 and 3. The 5-grams with a frequency of at least 3 occurrences, are shown for the Extravert group in Table 7. Introvert 5-grams failed to reach this frequency. For reference, the previous findings of the bigram analysis are also presented in Tables 4, 5 and 6. These represent the relative frequency ratio data and bigrams unique to Extraverts and Introverts respectively.

Trigram	Extr Freq	Intr Freq	Extr R. Freq	Intr R. Freq	Rel. F Ratio
a bit of	10	2	0.0007	0.0002	3.31
. i have	12	3	0.0009	0.0003	2.65
! ! !	17	6	0.0013	0.0007	1.88
. it was	19	7	0.0014	0.0008	1.80
. i think	8	4	0.0006	0.0004	1.33
. i am	13	7	0.0010	0.0008	1.23
. i was	9	6	0.0007	0.0007	0.99
for a bit	3	2	0.0002	0.0002	0.99
i am going	6	6	0.0004	0.0007	0.66
i have to	7	9	0.0005	0.0010	0.52
need to get	3	4	0.0002	0.0004	0.50
i'm going to	5	8	0.0004	0.0009	0.41
that i am	2	4	0.0001	0.0004	0.33

Table 1: Shared Extravert and Introvert trigrams

Trigram	Rank	-2logλ	Freq	Rel Freq
. . .	2	478.73	71	0.0053
looking forward to	5	328.73	15	0.0011
it was a	22	248.92	10	0.0007
really looking forward	19	270.95	6	0.0004
want to get	44	225.53	6	0.0004
next week .	47	222.33	6	0.0004
i will be	48	222.21	6	0.0004
going to get	9	306.28	5	0.0004
! it was	36	230.07	5	0.0004
i have been	37	228.85	5	0.0004
next week ,	38	228.05	5	0.0004
. . i	3	375.01	4	0.0003
! ! so	10	298.21	4	0.0003
the next week	20	267.20	4	0.0003
i'm looking forward	21	253.48	4	0.0003
a bit worried	23	247.71	4	0.0003
was a bit	26	242.06	4	0.0003
for next week	39	227.78	4	0.0003
going to do	18	273.76	3	0.0002
am looking forward	24	247.01	3	0.0002
will be able	28	240.29	3	0.0002
it was cool	32	234.94	3	0.0002
it was nice	34	233.68	3	0.0002
< END > next week	43	226.54	3	0.0002
and i am	46	223.07	3	0.0002
. . < END >	49	221.95	3	0.0002
it was really	50	221.69	3	0.0002
! ! not	11	290.85	2	0.0001
! ! it	12	289.13	2	0.0001
! ! ,	13	288.58	2	0.0001
! ! and	14	287.44	2	0.0001
! ! on	15	287.06	2	0.0001
so i am	29	240.01	2	0.0001
, it was	31	238.97	2	0.0001
i am looking	40	226.87	2	0.0001
but it was	41	226.82	2	0.0001
quite a bit	42	226.79	2	0.0001

Table 2: Trigrams unique to Extravert corpus.

Trigram	Rank	-2logλ	Freq	Rel Freq
going to the	13	188.42	7	0.0008
, but i	22	166.15	6	0.0007
i don't know	37	149.07	6	0.0007
going to be	10	197.73	5	0.0006
am going to	14	185.55	5	0.0006
. i don't	8	202.76	4	0.0004
managed to get	42	142.86	4	0.0004
in the evening	50	135.90	4	0.0004
going to see	2	239.05	3	0.0003
going to go	6	208.67	3	0.0003
. i got	24	163.75	3	0.0003
, and then	25	163.24	3	0.0003
. i played	26	162.53	3	0.0003
. i will	32	155.63	3	0.0003
. i wasn't	36	150.00	3	0.0003
, but it	41	145.19	3	0.0003
is a bit	45	137.72	3	0.0003
tomorrow i am	11	192.45	2	0.0002
i am not	16	183.39	2	0.0002
i am in	17	177.91	2	0.0002
probably going to	19	169.68	2	0.0002
it's going to	20	168.39	2	0.0002
going to book	21	167.60	2	0.0002
were going to	21	167.60	2	0.0002
, but it's	23	164.30	2	0.0002
trying to get	27	161.84	2	0.0002
be going to	28	161.26	2	0.0002
just going to	30	157.96	2	0.0002
was going to	31	155.84	2	0.0002
. i had	33	152.71	2	0.0002
went to see	34	152.68	2	0.0002
going to a	35	152.27	2	0.0002
, but that's	38	148.39	2	0.0002
, but there's	39	146.48	2	0.0002
, but he	40	146.29	2	0.0002
. i should	47	136.98	2	0.0002
. i still	48	136.57	2	0.0002
again . i	49	136.06	2	0.0002

Table 3: Trigrams unique to Introvert corpus.

Bigram	Extr Freq	Intr Freq	Extr R. Freq	Intr R. Freq	Rel. F Ratio
looking forward	15	4	0.0011	0.0005	2.49
it was	46	22	0.0034	0.0025	1.39
next week	24	12	0.0018	0.0013	1.33
a bit	29	15	0.0022	0.0017	1.28
up with	19	10	0.0014	0.0011	1.26
! !	45	24	0.0033	0.0027	1.24
will be	24	13	0.0018	0.0015	1.22
i was	33	18	0.0025	0.0020	1.22
at the	27	16	0.0020	0.0018	1.12
to see	32	19	0.0024	0.0021	1.12
which is	15	9	0.0011	0.0010	1.11
for a	34	21	0.0025	0.0024	1.07
i have	44	29	0.0033	0.0032	1.01
to get	34	23	0.0025	0.0026	0.98
. i	99	69	0.0074	0.0077	0.95
on friday	11	8	0.0008	0.0009	0.91
, and	48	36	0.0036	0.0040	0.88
and then	23	19	0.0017	0.0021	0.80
in the	41	34	0.0031	0.0038	0.80
apart from	6	5	0.0005	0.0006	0.80
i am	33	28	0.0025	0.0031	0.78
i think	16	14	0.0012	0.0016	0.76
, but	35	31	0.0026	0.0035	0.75
a lot	10	9	0.0007	0.0010	0.74
going to	36	33	0.0027	0.0037	0.72
a few	12	11	0.0009	0.0012	0.72
to do	23	23	0.0017	0.0026	0.66
i've been	9	12	0.0007	0.0013	0.50

Table 4: Shared Extravert and Introvert bigrams.

Bigram	Rank	-2 logλ	Freq	Rel Freq
. .	8	183.48	152	0.0113
of the	33	79.47	40	0.0030
, which	20	100.89	25	0.0019
had a	16	115.60	22	0.0016
which was	24	95.69	19	0.0014
new year	7	192.22	18	0.0013
got a	45	66.65	17	0.0013
a good	46	64.45	16	0.0012
forward to	26	94.76	15	0.0011
need to	28	89.99	15	0.0011
i'll be	22	98.70	14	0.0010
on saturday	27	90.94	13	0.0010
we went	42	67.54	11	0.0008
as well	43	67.18	11	0.0008
couple of	30	84.18	10	0.0007
want to	41	68.01	10	0.0007
the moment	44	67.09	10	0.0007
< END > hi	21	99.44	9	0.0007
able to	50	61.19	9	0.0007
take care	23	96.00	8	0.0006
catch up	39	70.50	7	0.0005
other than	49	62.84	6	0.0005

Table 5: Bigrams unique to Extravert corpus.

Bigram	Rank	-2 logλ	Freq	Rel Freq
· < END >	17	80.13	20	0.0022
i don't	18	78.77	18	0.0020
went to	25	63.53	15	0.0017
to go	34	56.65	14	0.0016
all the	47	43.06	12	0.0013
i went	50	42.70	12	0.0013
one of	32	57.45	11	0.0012
trying to	29	60.75	10	0.0011
i'm going	36	52.84	10	0.0011
i can	46	43.90	10	0.0011
on thursday	20	72.22	9	0.0010
don't know	21	69.76	9	0.0010
i've got	35	55.19	9	0.0010
lots of	26	62.29	8	0.0009
this week	39	48.51	8	0.0009
anyway ,	45	44.79	8	0.0009
should be	40	48.10	7	0.0008
on monday	41	47.91	6	0.0007
two weeks	31	58.65	5	0.0006
loads of	49	42.72	5	0.0006
< END > hello	44	45.05	4	0.0005
exam results	42	47.26	3	0.0003

Table 6: Bigrams unique to Introvert corpus.

5-gram	Freq	Rel Freq
. · · · it was	4	0.0003
really looking forward to seeing	3	0.0002
my plans for next week	3	0.0002
i'm really looking forward to	3	0.0002
· i'm looking forward to	3	0.0002
what i've been up to	3	0.0002

Table 7: 5-grams unique to Extravert corpus.

4. Discussion

Our discussion of these results will take the following form: Firstly we discuss the evidence from the trigrams and 5-grams which suggests different collocation usage by the Extravert and Introvert groups, and in particular whether a distinct pattern is present for the Extraverts; Secondly, we will evaluate the usefulness of the Extravert/Introvert characteristics summarised in Figure 1 which were formulated on the basis of the bigram data, and discuss whether they are supported in the current findings; Finally we assess the role of word collocation in personality language.

4.1 Extravert-Introvert collocations

The trigram analyses reveal an even more distinctive pattern of Extravert and Introvert language use, than was found for the bigrams. This is demonstrated firstly by the greater number of unique occurrences found for both personality types than was the case in the bigram analysis, and secondly by the non-significant correlation in the ordering of occurrence of trigrams shared between the two personality groups.

Turning to the 5-gram data, it can be seen that when a frequency cut-off of 3 occurrences is used, co-occurrence data is only found for the Extravert group. Given the modest data set, it is not surprising that few repeated 5-grams are found; indeed it could be argued that the relative difference in size between the Extravert and Introvert sub-corpora is responsible for this finding, although this in itself highlights the longer length of text produced by Extraverts, which is around 20% longer. However, when referring to data for 5-grams occurring with a frequency of 2, there are still disproportionately more of them for the Extravert group (n=56) than for the Introvert group (n=18). This pattern is also found from analysis of the whole of the trigram data occurring with a frequency of at least 2 and significance of $p < .005$. In this case, for the Extraverts 608 of 729 are unique, and for the Introverts this is 288 of 409, with 121 trigrams shared by both personality groups.

In order to better utilise the information that can potentially be provided by larger window *n*-gram analysis, a larger corpus would be preferable, along with a higher frequency cut off (eg. 5) and possibly also a statistical test of co-occurrence, like log-likelihood.

Before examining the trigram results in more detail, it is important that we clarify co-occurrence further. In the current analysis we have included or rather *not excluded* by way of stop list functors, punctuation, or rarer words and collocations, since the purpose of *n*-gram analysis in the current study is to find characteristic language patterns more generally, rather than the identification of, for example, key words.

We therefore distinguish co-occurrence more generally, into collocation, and colligation. Collocation, as we define it here, is what is perhaps more generally understood by the term *co-occurrence*, that is, 'the patterns of combinations of words (for example, with other words) in a text' (Oakes, 1998). Examples of collocation would be words which may occur separately, but occur together in a significant and meaningful way, in the way that *corpus linguistics* and *word frequency* may feature in the genre of corpus linguistics.

Colligation, on the other hand, is information which again is derived from co-occurrence information, eg. *n*-grams, but could not be described as collocation in the traditional sense. It is usually seen as more grammatically-oriented, covering the syntactic preferences of a word. For us, examples of colligation would be the positioning of words in relation to punctuation or other boundary markers, indicating that a particular word or token occurs in a text or sentence initial or final position. In the genre of formal letter writing, an example of a colligational co-occurrence might be "*start of document"* followed by *Dear*. Although punctuation is generally used to signal a sentence or phrase boundary, and is thus useful in determining colligation, we further distinguish between punctuation when used for a purely syntactic purpose, and when it is used to encode additional meaning, as is often the case in e-mails for example: multiple full stops or exclamation marks.

Given our hypothesis that Extraverts are more likely to use and re-use chunks of language, we would expect that collocations will constitute a larger proportion of total co-occurrences (and colligations a smaller proportion) for Extraverts, compared with Introverts.

Therefore in examining the co-occurrence data, we turn first to the trigrams which are shared by both Extraverts and Introverts. Here we can see that almost half of the trigrams contain elements of punctuation. Five of these provide colligations concerning (presumably) sentence initial constructions ([. *i have*], [. *it was*], [. *i think*], [. *i am*], and [. *i was*]), and appear to be favoured by the Extraverts. Note that [*! ! !*] is considered to be collocation, rather than colligation.

This use of colligation trigrams by Extraverts is perhaps unexpected. However, while the relative ratio suggests they are more characteristic of Extraverts, raw counts suggest they are used frequently by both Introverts and Extraverts. Indeed it may be the case that Introverts and Extraverts are using the same constructions differently. For example, examining the trigram data which is unique to the personality groups shows that whilst trigrams with the first element being a full stop are likely to indicate the end of a sentence for Introverts, for Extraverts this is more likely to be the last element of an elliptical [. . .].

Other patterns from the unique trigram data are that Extraverts show some use of colligation ([*next week .*], [*next week ,*], [< *END* > *next week*]). This seems to be largely topic specific, resulting from the extraposing of the author's current concern.

When this is contrasted with the colligation trigrams used uniquely by the Introverts, it can be seen that these contain a great deal more information about the relative focus and the syntactic constructions favoured by the Introvert authors. In choosing to write about their past or forthcoming week, rather than extraposing that time period, as in the case of the Extraverts, the colligations show that instead Introverts focus on themselves. Therefore a large proportion of their trigrams demonstrate a sentence initial first-personal singular pronoun, *I* ([. *i don't*], [. *i got*], [. *i played*], [. *i will*], [. *i wasn't*]). Furthermore, the colligation data of Introverts also demonstrate use of co-ordination, particularly *but* ([*, but i*], [*, but it*], [*, and then*]).

This data shows then, that Introverts do in fact show greater proportional use of colligation. We now turn to the collocation trigrams to examine the evidence for the frequent usage of chunks of text.

Both personality groups share the use of phrases such as *a bit* ([*a bit of*], [*for a bit*]) and *am going* ([*i am going*], [*i'm going to*]), although Extraverts prefer the former constructions and Introverts the latter. When the unique data for these personality groups is consulted, this pattern is borne out with Introverts' extensive use of collocations which include *going to* ([*going to the*], [*going to be*], [*am going to*], [*going to see*], [*going to go*]), versus those of the Extraverts ([*going to get*], [*going to do*]). Conversely, the Extraverts use more of *a bit* ([*a bit worried*], [*was a bit*]) versus the Introvert [*is a bit*].

Although featuring punctuation, [*! ! !*], is regarded as a collocation, and is a feature preferred by Extraverts. Examination of the unique data shows that this non-standard use of punctuation, along with the elliptical (...) are key features of Extravert texts ([. . .], [. . *i*], [*! ! so*], [. . < *END* >]).

The co-occurrences unique to Extraverts show a larger number of collocations. Some of these refer to the future, such as *will be* ([*i will be*], [*will be able*]), whereas the evaluative [*it was cool*], [*it was nice*] and [*! it was*] refer to the past. As previously mentioned, reference to the topic of *next week* occurs frequently ([*next week .*], [*next week ,*], [*the next week*], [*for next week*], [< *END* > *next week*]), as does *looking forward* ([*looking forward to*], [*really looking forward*], [*i'm looking forward*], [*am looking forward*]). These trigram patterns feature again

in the Extravert 5-grams in [. . . *it was*], [*really looking forward to seeing*], [*i'm really looking forward to*] and [*my plans for next week*].

On the basis of this evidence, it appears that Extravert and Introvert use of co-occurrence is different, with the Extraverts tending to use larger chunks of word collocations, and the higher proportion of Introvert colligations suggesting characteristic syntactic constructions. The co-occurrences which were shared by both groups were also shown to be used in significantly distinct ways.

4.2 Personality language style

Although the previous findings presented in Figure 1 were based upon bigram data, we now address whether the current extension of the analysis using higher *n*-grams still supports these broad personality language features.

Potentially using larger windows of text allows the identification of larger-scale features from the data, in the current case, patterns of between 3 and 5 words or characters. However, this also means that collocations of two words which co-occur with a large variety of words on either side will not show up in the current extension of the analysis. This means that whilst the Surface Realisation features (. . .) and (*!!!*) are very apparent in the trigram analysis, others such as the message initial *hi* or *hello* are not, since the name—or lack of name—which tends to follow is not a stable feature. Similarly, the bigrams characteristic of the Social Devices category *catch up* and *take care* also did not occur in the present analysis.

In a similar way, the result of analysis using 3-word windows on Message Planning and Expression features, is that the co-ordinations (*, and*) and (*, but*) are isolated in patterns which are even more strongly characteristic of Introversion (the previous bigram analysis found them used by both but preferred by Introverts). However, the Extravert feature (*, which*) was not found to occur in the present trigram or 5-gram analysis.

This pattern is repeated in the other bigram feature categories: for Quantification, Introverts do not demonstrate the large variety of features found originally, instead they make less use of (*a bit*) which is a rather vague, shared term used primarily by Extraverts; evidence of Modality is only found for the strongly predictive Extraverts (*will be*), but not for Introverts; the timid Ability of Introverts is found in (*trying to*) and the shared form (*going to*), but confident Extravert forms are not found.

Features expressing Valence were still found characteristically in the Extravert and Introvert texts: The former used expressions such as (*looking forward*) and *nice* and *cool*, with the latter employing the contracted negation *don't*. In the case of Self/Other Reference, although the Extravert tendency to refer to others was not maintained, further evidence for the mainly Introvert self-reference was found. Indeed, the colligations revealed interesting difference in the occurrence of the first-person singular, with Introverts tending to use this in the sentence initial position, whereas Extraverts were more likely to use it positioned within a sentence, or following elliptical (. . .).

These findings therefore largely support the previous Extravert-Introvert language features derived from the bigram analysis. Although the use of larger windows for co-occurrence analysis can uncover larger-scale language patterns, this can also result in the loss of patterns which only stably occur in two-word windows. Furthermore, the use of larger windows can

result in data sparsity, especially when using smaller corpora, and this is especially relevant for the Introvert data.

4.3 The lazy Extravert

In this paper we proposed that Extravert discourse strategy is based upon a kind of *laziness*, which manifests itself in their recycling of formulaic chunks of words. In our *n*-gram analyses we have demonstrated differences between Extravert and Introvert language usage which suggest that this is in fact the case. Note that we do not exclude the possibility that *everyone* re-cycles formulaic language, at least to some extent. The point is that Extraverts do so more than Introverts.

But why should Extraverts be particularly lazy and prone to re-using language features? Such behaviour is not without good reason since it serves the drives of the Extravert well. Earlier we described the Extravert as someone who is sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. Furthermore they think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic.

Through these personality descriptions we see an Extravert who wants to be the centre of attention, and as a result wants to gain the floor by quickly formulating a comment, or to hold on to it by continuing to talk. In contrast, the Introvert is less concerned with talking for talking's sake, but instead will be more inclined to enter into the conversation with a carefully considered contribution when they feel this is warranted.

These different conversational stances therefore impose a different set of constraints upon the Extravert and Introvert speakers. Introverts can afford greater mental resources in the planning and preparation of an utterance and thereby risk losing a conversational turn if another speaker formulates and executes a contribution more quickly thereby making the Introvert's irrelevant. Extraverts, when they are not speaking, are under pressure to quickly make a comment, thereby entering the conversation and gaining control of the floor. This process itself forms part of the Extravert's stimulation feedback loop, with fighting for the floor providing the stimulation which Extraverts crave.

We therefore propose that such pressure upon Extraverts to more quickly produce linguistic contributions leads to the employment of distinctive discourse strategies. Indeed, Furnham (1990) suggests that the Extravert has a more restricted code, which could well be the result of such constraints and would fit in with our observation of the reduced concreteness of such utterances, and the tendency to recycle pre-formed chunks of language.

Although previous discussion has concentrated upon the *spoken* language of Extraverts, we suggest that similar patterns occur in all naturalistic language production settings, since Extraversion is a stable trait which consistently influences an individual's behaviour. Instances where this may not play such a large role would be in carefully constructed written texts and where several iterations of editing are likely to occur. Given that the style of e-mail is considered to be close to that of oral communication (Bälter, 1998), we would expect that laziness, typical of Extraverts, is found in e-mails and similar texts.

5. Conclusion

We have shown that Extraverts and Introverts use larger-scale co-occurrences of words in characteristically distinct ways through *n*-gram analysis. This has extended recent work which derived Extravert and Introvert linguistic behaviour using a combination of techniques from psycholinguistics and statistical natural language processing.

This differentiation between personality groups lends support to our hypothesis, based on previous findings, that Extraverts are "lazy" and use larger-scale collocations of words in order to spend less time planning discourse. A greater proportion of co-occurrence information for Introverts was colligational and related to the structure of their text. Our trigram and 5-gram analyses broadly support previous findings for bigrams. However, we note that in some cases bigrams may be more informative, and that care should be taken with regard to data-sparsity with larger *n*-gram analyses.

Further, more technically sophisticated analyses can be carried out: we envisage the use of machine learning techniques to automatically classify texts on the basis of the distinctive features we are isolating, along with further *n*-gram analysis exploiting 'parts of speech' tags.

6. Acknowledgements

Thanks to Elizabeth Austin, James Curran and our anonymous reviewers for helpful advice and comments. This work was supported by the Economic and Social Research Council (Award R00429934162), and the School of Informatics, University of Edinburgh.

References

Bälter, O. (1998). *Electronic Mail in a Working Context*. Ph. D. thesis, Royal Institute of Technology, Stockholm.

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology*, **4**, 1–7.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, **29**, 433–448.

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

Dewaele, J.-M. and Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning*, **52**, 265–324.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**, 61–74.

Eysenck, H. and Eysenck, S. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself, but Neuroticism is more of a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Boston, MA, August 2003.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp363-368. Fairfax VA, August 2002.

Jonassen, D. and Grabowski, B. (1993). *Handbook of Individual Differences, Learning and Instruction*. Laurence Erlbaum Associates, Hillsdale, NJ.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

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