Original citation:

Wharton, S. (2012). Epistemological and interpersonal stance in a data description task: findings from a discipline-specific learner corpus. English for Specific Purposes, 31(4), pp. 261-270.

Permanent WRAP url:
http://wrap.warwick.ac.uk/50365

Copyright and reuse:

Publisher's statement:

A note on versions:

For more   information,
please contact the   WRAP Team at:

warwickpublicationswrap

highlight your research

http://go.warwick.ac.uk/lib-publications

Accepted version

Epistemological and interpersonal stance in a data description task: findings from a

discipline-specific learner corpus

1.   Introduction

This paper describes and interprets findings from a small corpus of NNES undergraduate writing

in the discipline of Statistics. Analysis focuses on a specific rhetorical task, that of data

description. It examines both the content of the student writing, and the range of resources

deployed by writers to vary their stance towards statistical propositions. In the light of an

interview with a specialist informant, the paper also discusses pedagogic uses of the learner

corpus and the analytical findings.

Data is drawn from a first year Statistics course at Warwick University, entitled "Stats Lab". A

task from this course presents students with a table containing data collected by the

Organisation for Economic Cooperation and Development (OECD) on total health expenditure

per capita in US dollars and life expectancy at birth for 32 OECD member nations. Students are

then asked to undertake three activities: to plot a graph of life expectancy vs. total health

expenditure, to discuss what the graph shows, and to use the graph to estimate a life

expectancy for Chile. The task therefore involves a combination of mathematical reasoning and

written expression.

The student answer data is of interest for a number of reasons. Firstly, and despite the fact that these disciplines attract high numbers of NNES students, student writing in Science and Technology is relatively under researched. This is perhaps due to a perception that rhetorical acts of description and interpretation are less complex in hard science writing than in social science writing (Hyland, 2006). A recent article in a Statistics Education journal (Parke, 2008) and a presentation at a Statistics Education conference (Forster et al., 2005) both highlight the need to teach undergraduate Statistics students to write about Statistics effectively – to communicate through language as well as through mathematics – and argue that this is an under-researched area.

Then secondly, research into academic text writing does not often focus on data description sections. More work has been done on sections such as introductions or conclusions which may be seen as rhetorically more varied and therefore more challenging for students. And yet, data description is a frequently set task: Braine (1989) surveyed typical non-exam writing tasks in undergraduate Science and Technology, and found that all categories of task except one "require the summary of information obtained from other sources or from the student's own observation" (p13). More recently, the distribution of university assignments found in the British Academic Written English Corpus (BAWE) suggests that data description assignments are relatively frequent in Science and Mathematics courses. BAWE categorises data description assignments under the genre family label "Exercise" (BAWE corpus manual, p47). BAWE does not identify Statistics as a separate discipline, but categorises assignments produced for a

Statistics course under the discipline of Mathematics. Of the 34 Mathematics assignments in BAWE, 15 are "Exercises".

Research reports by Statistics educators also suggest that data description tasks are frequently set in university contexts and are experienced as challenging. (Forster et al., 2005; Lipson & Kokonis, 2005; Parke, 2008). MacGillivray (2005) comments: "Learning to communicate *about* graphs, plots and tables can be challenging, as it is a highly interdependent process combining confidence and awareness of (i) the scope and limitations of the various representations, (ii) discerning key features midst variation, and (iii) synopsis and efficient use of language". (2005, p.3).

2.   Research focus: writer stance towards propositions

The term *stance* refers to the ways in which authors project themselves into texts, often evaluatively, often to indicate their relationship and attitude to their subject matter and to their readers. It is concerned with the epistemic domain (stances as to truth, certainty), the attitudinal domain (feelings and value judgements), and with dialogic space (the extent to which authors acknowledge the possibility of alternative positions). In spoken text, stance is constantly adjusted in interaction with the interlocutor. In written text, stance is constantly adjusted in interaction with the construed readership.

The language of stance has been researched from a number of perspectives including Appraisal, Attitude, Epistemic Modality, Metadiscourse, and Voice. There are important differences in the

ways that researchers in these different traditions understand the scope of stance, and each

tradition has developed its own range of taxonomies for describing stance resources. This

makes it difficult to directly compare research undertaken from the different perspectives.

Nevertheless, there are enough similarities between the different conceptualisations of stance

to make it possible to relate research findings emerging from the different traditions.

Essentially, all frameworks relate stance to "the speaker's opinion or attitude towards a

proposition that the sentence expresses or the situation that the proposition describes" (Lyons,

(1977, p.452).

The importance of stance in academic writing has been thoroughly discussed, often from a

genre or discourse community perspective. As Hyland (2005, p. 41) argues, the expression of

varying stances towards propositions allows writers not only to position themselves vis à vis the

information that they are discussing, but also to position themselves vis à vis a community of

readers. By expressing a range of degrees of commitment towards propositions, writers can

express their views with delicacy at the same time as showing sensitivity to community

expectations. For example, an appropriate deployment of stance resources could indicate

knowledge of what sort of proposition may consensually be seen as likely to be definite, and by

contrast where a definite commitment would be seen as inappropriate. Koutsantoni (2004),

Bruce (2009) and McGrath & Kurteeva (2011) all include some notion of stance in their analyses

of expert writing in a range of disciplines; all argue that an appropriate deployment of stance

resources is one of the ways in which a writer indicates their membership of a disciplinary

community. For example, Bruce (2009), who looks at results sections of Sociology and Organic

Chemistry articles, quotes an informant from Sociology as proposing that "the reporting of research findings has to acknowledge multiple views of reality", and an informant from Organic Chemistry as suggesting that "key values are the need for preciseness of detail as well as conciseness". (2009, p. 111).

And yet, various researchers have shown that NNES student writers experience difficulty in deploying the linguistic resources necessary to express a nuanced range of commitments to propositions. Chen (2010) compared the writing of college level Chinese students with comparable texts from the British National Corpus and discovered that the Chinese writers use epistemic modality far less frequently than the BNC texts. She also found that they tend to over-use strong, or boosted, assertions. [Author] (in review) finds that Chinese writers studying through English in China make more use of strong assertions than do Chinese writers studying through English in the UK. Hyland & Milton (1997) suggest that the acquisition of epistemic modality is difficult due to diversity of language form and the contextual determination of the function of such forms, as well as due to cultural differences about appropriate uses of modality. Koutsantoni (2006) argues that students find it difficult to project an appropriate level of certainty and authority in their texts because they are inhibited by their perceptions of power relations between themselves and their readers.

A factor influencing this difficulty may be students' relative lack of exposure to a range of stance resources. Chen (2010) and Vellenga (2004) each reviews the treatment of epistemic modality in a range of writing textbooks. Both studies argue that the treatment is lacking or

5

misrepresentative in some books, perhaps adding to the difficulties of students. Biber (2006) in a survey of texts from a broad range of university disciplines, notes the overall rarity of stance expressions in university textbooks.

My decision to focus on the language of stance towards propositions emerged both from a bottom-up coding of texts (see below) and from an awareness, indicated above, that the issue is considered significant in EAP research and teaching.

3.   Methodology: Research corpus and analytical framework

Writing was collected during November 2010 from a single cohort of first year undergraduate Statistics students. Students who consented to participate provided copies of their hand-written assignments, which were transcribed and saved in .txt format. The corpus of task answers used for this research consisted of 40 student texts, a total of 4705 words excluding formulae, graphs etc. The shortest text was 47 words and the longest 275 words.

To explore the nature of the writing, I used a bottom up approach to qualitative content analysis (Hsieh & Shannon, 2005; Zhang & Wildermuth, 2009). Qualitative content analysis is a systematic approach for looking at patterns of content and/or expression in text. When used inductively as here, it proceeds via examination and re-examination of texts, searching for categories of meaning which the texts may have in common (Richards, 2003) .

I began with open coding of the 40 texts, using the CAQDAS programme *Nvivo 8*. Like other

programmes for computer assisted qualitative data analysis, Nvivo is designed to facilitate the

organization and analysis of non-numerical, unstructured data (Seror, 2005; Schönfelder, 2011).

It allows users to classify and arrange information in different ways, and thus to search for

relationships in the data. The task of qualitative content analysis is thus made more

manageable.

Examining my corpus, I saw some potential similarities across texts, based both on the subject

matter of the texts and on the language in which writers had expressed themselves. This initial,

open coding led to a number of potential nodes to describe such similarities. Through

examination and re-examination of the texts, and consideration of the patterns which seemed

to recur most frequently, I arrived at an apparently stable range of nodes. I was then able to

group these into two major themes: *Common Content Assertions* and *Common Stances in*

Assertions. I will now explain each theme in turn.

The first theme, *Common Content Assertions,* was developed through my observation that

certain propositions, though expressed through a variety of language resources, occurred

frequently across the texts. I found five content propositions which occurred frequently (a

minimum of 25 instances in this data), and no other content propositions with more than 5

instances. I therefore treated the five frequent propositions as nodes for coding. They reflect

not only the extent to which students had addressed all aspects of the prompt but also,

importantly, students' interpretation of the instruction to 'discuss what the graph shows'.

Together, they represent some kind of consensus as to what the answer might be expected to include.

The following diagram, then, illustrates the five nodes of content proposition which I grouped together to form the larger theme entitled *Common Content Assertions*. The node labels will be explained in more detail in the Results section of the paper.
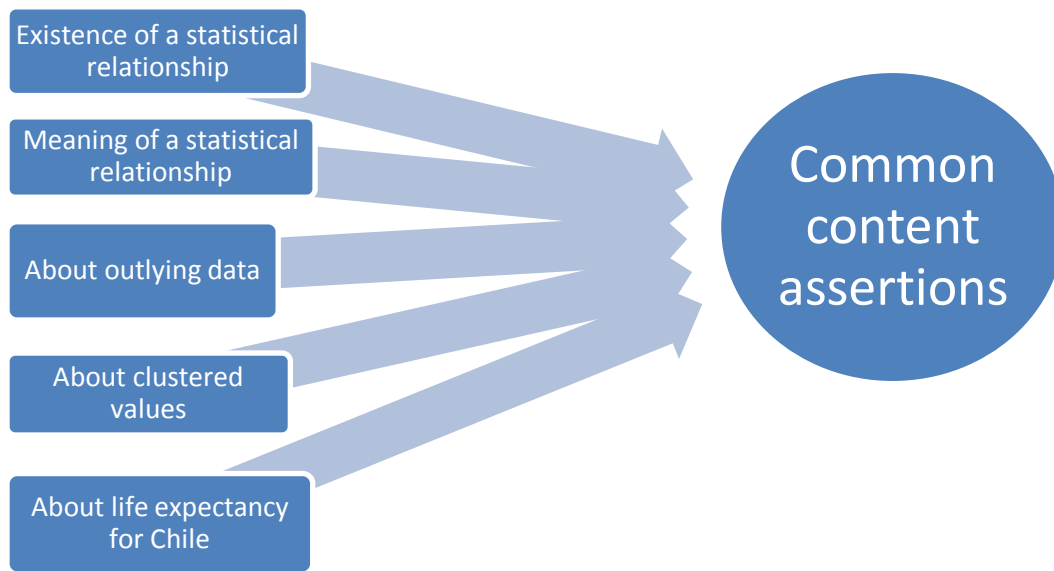
Illustration 1: Categories of content assertions

The second theme, *Common Stances in Assertions* arose from observations about different positions that writers appeared to take up vis à vis content propositions. Again, open coding of instances where writers appeared to express a stance enabled me to identify categories of

stance which were frequently used. I coded stance expressions appearing in any proposition, not only those which could be grouped into the 'frequent' content categories shown above. This meant that the picture of stance which emerged was not limited to the 'common consensus' content of the texts.

The following diagram illustrates the categories of stance type which I grouped together to form the larger theme Common Stances in Assertions. This theme reflects the observation that writers evidenced different ways of positioning themselves vis à vis the ideas in their texts. Again, the node labels will be explained in more detail in the Results section of the paper.

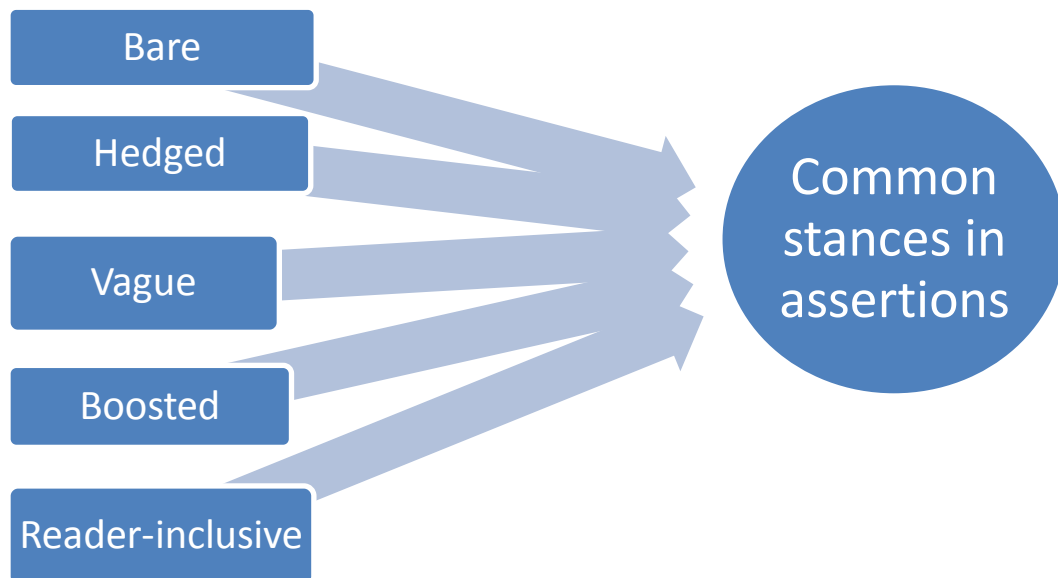**Illustration 2: Categories of stance in assertions**

Having arrived at a node structure which I considered stable, I used it to re-code the data and obtain quantitative results. When re-coding to content nodes, I coded only those propositions falling under the categories which I had found to be frequent; but when re-coding to stance nodes, I coded stance expressions appearing in any proposition, not only those coded to the 'frequent' content nodes. Where necessary, propositions were coded for more than one stance type; for example, 'The life expectancy for Chile could be (hedged) around (vague) 74.'

I coded main clauses and paratactically linked sub-clauses individually. I did not apply separate codes to hypotactically linked sub-clauses or to rankshifted clauses functioning as subjects or complements. These latter clause types do not form main propositions in this data, and to code them separately would have distorted any frequency counts. So for example in the extract 'the other data tend to form a positive correlation, indicating that higher total health expenditure per capita leads to a higher life expectancy at birth.', 'tend to' and 'indicating' were both coded, but 'leads to' was not.

To make these explanations clearer, I will provide an example of an answer as re-coded at both content and stance nodes. Propositions which are not shown as coded for content were not considered to belong to any of the common content nodes.

> The graph *shows (bare)* weak positive correlation [EXISTENCE OF STATISTICAL RELATIONSHIP] which *indicates (hedged)* the relationship that the longer life

10

expectancy at birth the higher total health expenditure. [MEANING OF STATISTICAL RELATIONSHIP] It *can be observed (bare)* that United States is an outlier. [STATEMENT RE OUTLIER] The life expectances at birth is 77.7 in the US and $6931 in total health expenditure per capita, which is *almost two to three times (vague)* than other countries in the similar life expectancy range (76-79). The outlier also *suggests (hedged)* that the median of life expectancy at birth and total health expenditure is a more appropriate summary for the OECD members than the mean. *About 1/3 (vague)* countries are in the 73-78 life expectancy range whereas *other 2/3 (vague)* countries lie on the range of 78-73. [STATEMENT RE CLUSTERED DATA] All countries except US, the outlier, *have (bare)* the health expenditure per capita lower than 5000 dollars. This figure of 23 countries is *around 2000-5000, (vague)* while that of only 9 countries is *below 2000 (vague)*. The best fit line is drawn on the graph above, the equation for the line *is (bare)* <FORMULA> By substituting <FORMULA> the life expectancy at birth on Chile *was (bare)* 71.5 years. [ESTIMATE FOR CHILE].

The node structures illustrated above are the result of bottom up coding, which means that the categories are not necessarily related to any pre-existing framework for analysis. In this, my approach is different from researchers such as Koutsantoni (2004), Wu (2007) or McGrath & Kurteeva (2011) who all comment that they used frameworks developed by previous researchers as a starting point to code their own data. There are, of course, advantages and

disadvantages to each approach. Beginning with a pre-existing framework allows the researcher to work with a larger set of data, and also allows the researcher to investigate the extent to which the chosen framework, usually well known, is successful in accounting for the particular data set under investigation. However, I would argue that a bottom up approach is more rewarding to use with a small set of data drawn from a genre which has not been extensively researched. Specifically, it allows categories to emerge with lines of similarity and difference which may not be reflected by any given pre-existing framework, because they may be very specific to the data being discussed. Importantly, a bottom up approach does not preclude reference to a well known analytical framework at a later stage of data examination – as I will discuss below, I myself make use of the Engagement sub-system of the Appraisal framework (Martin & White, 2005) to better interpret how the categories which I identified seemed to be functioning in these texts.

4.   Results

4.1   Common content statements across texts

As was indicated above, there were certain propositions – expressed in whatever language form – which appeared in many of the texts. This was the case despite the considerable variation between the texts in terms of length and level of detail, and it seems likely that it can be attributed to the specificity of the task. On the other hand, given this specificity, it is perhaps noteworthy that there was any variation at all. The following table shows the frequency of occurrence of the content propositions:

Content Proposition	Number of instances in data	Number of texts (of 40) in which instances occurred
A statement of a statistical relationship between variables on the graph	39	39
A statement interpreting the meaning of this relationship	24	21
A statement about clustered values	25	23
A statement about an outlier value	29	28
An estimate of the life expectancy for Chile	30	30

Table 1 Frequency of occurrence of content propositions

The prevalence of five common content categories indicates a high level of agreement between writers over how to interpret the instruction to 'comment on what the graph shows'. 39 of 40 writers describe the statistical relationship shown by the data, using such terms as 'weak positive correlation' or 'non linear relationship'. 21 writers add a statement which hypothesizes as to the interdependence of the two variables and/or as to the direction of causality, e.g. 'higher health expenditure may lead to higher life expectancy.' Some writers have also chosen to comment about the ways the data clusters and about any values that do not fit the overall pattern, with 23 texts including at least one reference to clustering of data (e.g. 'Most of the data is grouped in one cluster.') and 28 texts including at least one reference to an atypical,

outlying value (e.g. 'It is clear that United States is an outlier in the graph'). However, the results also show that not all students have engaged with all aspects of the task. The task brief specifically requests an estimate of the life expectancy for Chile, yet only 30 of 40 writers have provided it.

It could be argued that these common content assertions have the status of expected 'moves' in these texts. The term 'moves' usually refers, of course, to illocutionarily motivated stretches of text within a particular genre (Swales, 1990). Here, the 'moves' are motivated by something much more concrete than a genre, i.e. the details of the task brief and the very definite expectations of the pedagogic setting.

4.2   Stances in Assertions

As was discussed above, a process of bottom up coding using Nvivo 8 led to 5 categories of Stance in Assertions: Bare, hedged, vague, boosted, and reader-inclusive. The table below shows their frequencies in the data.

Stance type	Number of instances in data	Number of texts (of 40) in which instances occurred
Bare	119	39
Hedged	81	32
Vague	44	23
Boosted	14	12
Reader-inclusive	15	14

Table 2 Frequency of occurrence of stance types

These categories, and the specific language resources found in the statements which are grouped in the various categories, can be usefully discussed with reference to relevant elements of Martin & White's (2005) Appraisal Framework, which 'is concerned with the interpersonal in language, with the subjective presence of writers/ speakers in texts as they adopt stances towards both the material they present and those with whom they communicate' (2005, p.1).

This framework acknowledges that writers express different degrees and kinds of commitment to propositions for both epistemological and social reasons. Epistemological reasons relate to uncertainty regarding content; writers need to be accurate and to be cautious about what they do not know. Social reasons relate writers' awareness that their position is one of many possible ones, and to the extent to which they choose to leave room in the text for alternative positions or voices. Koutsantoni (2006) argues that this choice is particularly complex in genres where writers are positioned as having a lower status than readers, as in the student texts examined here.

Martin & White (2005, p. 34 ff) describe the appraisal framework as containing three interacting domains: *Attitude*, concerned with feelings and emotional reactions and judgements; *Engagement*, concerned with the sourcing of attitudes and the acknowledgement of different voices in discourse; and *Graduation*, concerned with grading phenomena, showing

how strong or weak a stance may be. To discuss my data I draw particularly on the domain of *Engagement*, which seems to be the most relevant to these texts, where students may or may not choose to acknowledge possible responses to the propositions advanced in their texts. In the next section, I will discuss the ways in which each stance category can modify an assertion and open up, or close down, dialogic space.

4.2.1   Bare stances

Bare-stance assertions are definite propositions, expressed without any language to suggest doubt on the part of the writer or, conversely, any language to emphasise the writer's certainty. Martin & White (2005) categorise such assertions as monoglossic: they make no reference to other voices and do not recognize alternative positions. This was the category with the most instances, with bare assertions appearing 119 times over 39 texts.

Examining the language chosen by writers to make bare assertions, an obvious pattern emerges. By far the most frequent word chosen is simply <u>is</u>, appearing 57 times including lemma forms. Examples are 'there is an outlier', 'there are a few outliers'. <u>Have</u> is also frequent (13 times). (e.g. 'it has the weak positive correlation'). An important word is <u>show</u> , appearing 29 times including lemma forms. (e.g. 'the scatter plot above shows that….') Previous research using a stance approach (Koutsantoni 2004) categorises the word <u>show </u>as expressing certainty and commitment. Her corpus includes instances of <u>show</u> being used to imply that the author concurs with overall research findings, e.g. in the phrase 'the performance of Gas has been shown to be better than that of' (2004: 173). However, in my data, the word <u>show</u> is used to

describe what is seen on a graph. It therefore seems more appropriate to classify it as a bare assertion for this corpus.

Martin & White (2005, pp. 98-100) argue that a frequent rhetorical effect of the bare assertion is to seem to take for granted that a reader can share a writer's point of view. Bare assertions construe a putative reader who finds the propositions so expressed to be unproblematic. I will examine below which kinds of proposition tend to receive such treatment in this writing.

4.2.2   Hedged stances

These are where the writer indicates some hesitancy about the content of the assertion. For Martin & White, such a stance is heteroglossic; the hesitancy expressed allows the authorial voice to entertain the possibility of other views.

In this corpus such statements appear 81 times over 32 texts. The most frequent hedging technique seems to be modality, or more specifically modalisation, the use of modality "to argue about the probability or frequency of propositions" Eggins (1994; 179). The most frequent hedging item (11 instances) is may: 'life expectancy value for Chile may be near 73.3 years'. Could appears 5 times (e.g. 'this could be due to'). Turning now to lexical hedges, indicate as an alternative to show appears 6 times, (e.g. The graph also indicates). Estimate appears 16 times, but it is difficult to know what significance to attach to this as the term is given in the task brief and students may simply be repeating it. Other items used as hedges included possible (4 instances) and relatively (also 4), maybe (1 instance) and might (2

instances). All of these items have what Martin & White (2005) would term a dialogically expansive function; they indicate the author's openness to other possible voices and views.

4.2.3   Vague stances

Stretches of text coded to this node contain what Chanell (1994) terms *vague quantifiers*, or more specifically *approximators:* phrases such as 'most of the points' or 'over 80'. Expressions like this are not hedges in the sense discussed above, in that they do not include explicit reference to the stance of the authorial voice. Rather, they "introduce fuzziness *within* the proposition that the speaker is expressing" (p17, emphasis in original). In this corpus, the inclusion of vague quantifiers in an assertion about data enables the writer to avoid committing to a very precise assertion. It seems likely that this technique indicates an epistemically motivated hedge, as distinct from the dialogically expansive hedges discussed in the previous section.   Its presence suggests that the student writers have some awareness of levels of accuracy required in the discipline. The technique is frequent, being identified in 44 instances over 23 texts. The most frequently occurring vague quantifier is <u>about</u>, occurring 14 times (e.g. From the graph you can see the life expectancy in Chile is about 73 years old; about 50% of the countries have life expectancies ranging from 79 to 82 years and spend about 2500 to 3700 USD per capita on health).

Other items used are <u>between</u> (7), <u>most</u> (7), <u>half</u> (5), <u>more</u> (4) and <u>over</u>(4). There are no other vague quantifiers which appear more than 3 times. Those used with lower frequency include <u>approximately</u> (2) and <u>around</u> (2). There seems then to be a preference for over quantifying

rather than underquantifying – <u>most</u>, <u>more</u> and <u>over</u> account for 15 occurrences between them. I found no examples of <u>few</u>, <u>less</u>, <u>fewer</u>, or <u>under</u>. <u>Below</u> appears once.

<u>4.2.4   Boosted stances</u>

These are assertions where the writer marks extra certainty regarding the proposition, often by using items such as <u>clearly</u> or <u>obviously</u>. Martin & White (2005, pp. 98, 121-127) might see these utterances as fulfilling a *proclaim* function. They are heteroglossic, in that they acknowledge the possibility of alternative points of view; but unlike the hedges discussed above, they result in dialogic contraction rather than expansion. The textual voice "sets itself against, suppresses or rules out alternative positions" (2005: 98).

In this corpus, there are far fewer boosted assertions than hedged assertions. The most frequent items used are: *clear* & lemmas (5 instances, e.g. it has shown a clear positive correlation), *easily* & lemmas (3 instances, e.g. we can easily see that…) and *obviously* (2 instances e.g. Obviously it is not a reliable data…)). For Martin & White, the key difference between these assertions and the bare assertions discussed above is that, while both sets construe a reader who agrees with the writer, assertions fulfilling a *proclaim* function also acknowledge the existence of other readers who might not agree. That is why they boost the certainty of the assertion.

<u>4.2.5   Reader-inclusive stances</u>

Stretches of text coded to this node contain inclusive <u>we</u>. Inclusive <u>we</u> is often seen as an "engagement marker" (Hyland, 2005; 151) which functions to bring the reader into the discourse. In academic writing, for example, it may serve to claim that writer and readers are members of the same disciplinary community. In this corpus, it seems to have a slightly different function; to construct the reader as agreeing with the writer. In this data, inclusive <u>we</u> is associated overwhelmingly with the metaphorical use of the verb <u>see</u> – 9 of the instances are some variation on the phrase 'we can see'. This phrase expresses the alignment of the writer and the putative reader with the assertion (Flottum et al., 2006). Luzon (2009), also investigating a learner corpus, notes instances of <u>we</u> + <u>see</u> used to emphasise a claim, and argues that this pattern would be absent in expert writing. In Martin & White's (2005) terms, this pattern may be a form of dialogic contraction, explicitly representing writer and readers as interpreting data in the same way.

For these reasons, my first thoughts were to interpret statements under this category as having a *proclaim* function, similar to boosted assertions. However, as I will discuss below, my specialist informant had a different view.

4.3   Relationships between content statements and stances

Having explored the data from the perspective of frequent content statements and frequent stances, I next looked at whether there was any relationship between type of content statement, and type of stance; whether writers in this corpus showed any tendency to take common stances on certain content statements.

Quantitative results can be shown in the following table:

	Bare stances	Hedged stances	Vague stances	Boosted stances	Reader-inclusive stances	Total stances expressed in each content category
Existence of statistical relationship	31	4	0	3	3	41
Meaning of statistical relationship	7	15	0	1	1	24
Clustered values	6	3	17	0	2	28
Outlier value	15	8	0	6	2	31
Life expectancy for Chile	3	21	8	0	0	32
Total instances of each stance type in all content categories	62	51	25	10	8	

Table 3 Frequency of stance types in frequent content assertions

Readers will observe that the figures given for total stances in each content category are higher than the figures given for instances of propositions in that content category in table 1 above. This is because some propositions included more than one stance expression. Similarly, readers will observe that figures for the total instances of each stance type are lower than those given

in table 2 above. This is because table 2 includes stances taken in all propositions in this data, whereas this table includes only stances taken in propositions in the five highlighted content categories.

From table 3, we can see some possible patterns regarding the stance types which student writers tend to choose when expressing different propositions. For example, bare assertions, which are frequent overall, appear much more frequently in assertions about the existence of a statistical relationship than in any other type of assertion. Boosted assertions are far less frequent overall, but in this data are strongly associated with statements about an outlier. Patterns which can be observed from the table begin to suggest a picture of these writers' views about what sort of proposition calls for what sort of stance. Their decisions are no doubt partially motivated by the extent of their epistemic certainty but, as Koutsantoni (2006) has argued, they are also likely to be motivated by their understanding of their audience's view about the level of certainty which it is appropriate for them to express. In the next section, I will discuss the appropriacy of the choices they have made in the light of an interview with a specialist informant.

5.   Discussion: the perspective of a specialist informant

I conducted an interview with the Statistics Department lecturer with overall responsibility for the course on which this writing was produced. Our discussion focused on the ways in which the students' deployment of stance resources was, or was not, appropriate. My informant emphasized firstly that the task analysed was drawn from an *applied* Statistics course, where

the students were asked to work with authentic, external data; on other courses, "data" might legitimately be invented in order to demonstrate mathematical procedures. He emphasized that authentic data rarely conformed precisely to statistical models, and that this in itself meant that it was important not to make categorical claims about it.

I asked for his views on the common content categories which students seemed to have chosen and whether these were appropriate. My informant said that this would very much depend on the graph that students had managed to draw; different mathematical choices may have led them to different shapes of graphs and distributions of points. He felt that statements about the existence of a statistical relationship, and about possible clustering and outliers, would however be very likely given the data. He commented that a statement about the possible meaning of a statistical relationship, e.g. a direction of causality, was not at all called for by the question and could not be legitimately inferred from any graph the students would draw.

Overall, my informant felt that the different content categories which had been generated by the task would require very different stances towards the assertions. I will therefore discuss each content category in turn.

The most frequent content statement, about a relationship between variables, was expressed most frequently through a bare assertion. Some examples are 'The graph shows a weak positive correlation', 'There is a weak positive correlation between the two variables'. A limited number of writers boosted the statement in some way (e.g. The graph shows a clear

relationship between health expenditure and life expectancy), and a very few hedged their

assertion (e.g. Generally, it shows a very weak linear correlation...). My informant commented

that statements about the existence of the statistical relationship were relatively

uncontroversial, so that a bare assertion – the most popular choice – was appropriate.

The next most frequent content statement was the estimate of life expectancy for Chile. Here,

we see that the expression of some level of hesitancy was frequent, with 21 hedged assertions.

As we said above, this result should be interpreted with some caution, since the word estimate,

was coded as a hedge even though it appeared in the assignment brief. Examples of hedges

statements using the word estimate included 'The estimated life expectancy at birth of Chile is

73.4 according to the line drawn; Through the graph, a life expectancy of 75 years is estimated

for Chile.' Assertions using other hedging techniques included: 'Therefore, the life expectancy

value for Chile may be near 73.3 years.; The life expectancy value for Chile could be around 74'.

There were 8 instances of vague quantification, sometimes appearing in assertions which were

already hedged, as in the previous example. Only 3 writers used a bare assertion here, and

none used a boosted assertion.

Vagueness about life expectancy – whether expressed through vague quantifiers or hedges -

drew particular interest from my informant. He commented that students seemed to be

hesitating about the precise value to suggest (using phrases such as 'around 74' or 'between 73

and 76') whereas it would be more appropriate to read a precise value off the graph they had

drawn, but to express some hesitancy about how sure they could be that the value was correct.

In other words, he felt that writers should attribute any uncertainty to the validity of the graph, rather than offering an imprecise result. It remains a matter of speculation whether the students who used the word 'estimate', eg 'life expectancy for Chile is estimated to be 72.8 years' were attempting to do as he would advocate, or whether they were unconsciously taking the word from the prompt.

The third content category, statements about outlying data, exhibits one of the least clear patterns, with 8 being hedged, (e.g. 'we may regard it as outlier'; 'The graph also indicates an extreme value which may possibly be an outlier') and 6 being boosted ('There is an obvious outlier, the United States'; 'It is clear that United States is an outlier in the graph') and 15 being expressed through a bare assertion ('There is an outlier which is US'; 'One outlier is identified').

My informant particularly highlighted the outlier category, as one where more students seemed to have made inappropriate stance choices. He noted the comparatively high number of bare assertions and boosted assertions. He explained that it was not possible from the data given to establish that any particular value was an outlier, it would only be possible to say that it looked as if it might be. He further explained that this was a high stakes decision: once a value has been categorised as an outlier, it is not included in any subsequent calculations. He would have expected students to hesitate before classifying a value as an outlier, and ideally to comment on the consequences of excluding or including it in calculations.

The fourth content category, statements about clustered data, account for by far the most instances of vague quantifiers in assertions: (e.g. 'Over half of the data is clustered around the 79-82 region'; 'There are also clusters of data at high life expectancy at birth and total health expenditure per capita'; 'More than half of the countries have life expectancy over 78'). This usage was seen as appropriate by my informant. He pointed out that at this stage of their studies, writers were being asked to look at dispersions of points around graphs and make qualitative, exploratory observations about any potential patterns. Statements such as 'half of data concentrate around' are therefore very appropriate.

This leaves the fifth category, statements about the meaning of the relationship between the variables. As was noted above, my informant commented that such a statement is not fully warranted by the data given for this task, and he was relieved to observe that 15 of the 24 students who had included such a statement, had hedged it. (e.g. 'A higher total health expenditure per capita *may* result in a higher life expectancy at birth'; '*In general*, it can be said that if a country has higher health expenditure per capita, the life expectancy of the citizen will be relatively higher'). 7 statements however are in the form of bare assertions, (e.g. 'This means that spending more expenditure on health will lead an increase in life expectancy'; . 'total health expenditure increases as life expectancy increasing'), so this is by no means a universal decision.

Reader-inclusive stances are reasonably evenly distributed among the content categories. As I commented above, I first interpreted this stance as having some sort of boosting function. My informant, however, suggested that this use by students may be an inappropriate transfer of

language used in their lectures. He suggested that phrases like 'we can see' would typically be used by a teacher commenting on a graph or other data projected on a screen. In such a context, the use of the phrase would indeed seem to have an engagement function, with the speaker including the student listeners in the interpretation of the data under discussion. In a student answer, however, this pragmatic purpose was no longer appropriate.

6.   Pedagogic implications

Work by Statistics educators reviewed in this paper tends to look at Statistics writing from a macro-organisational and content based perspective: showing students what should be included in a section of e.g. a written report or a summary of results. For example Parke (2008) argues that to understand what is expected in terms of content is an important part of learning how to tackle a fairly general prompt to a Statistics writing task. She gives details of a pedagogic intervention in which students worked in groups on sample answers in order to decide which content propositions should be included, and thereby 'become familiar with the elements that constitute a complete and correct interpretation' (p 8/21).

My work also indicates that students need guidance as to appropriate content. The analysis of content categories, taken together with the informant interview, was sufficient to show what content might be expected, but it also showed that not all students had included all necessary content. Pedagogic work along the lines described by Parke (2008), discussed above, may well be of use to these students too.

My research also suggests that students need guidance as to the appropriate stance to express vis à vis statistical propositions. MacGillivray (2005, p3) quoted above, emphasized that data description tasks are challenging in part because they require students to show awareness of the scope and limitations of the data they see, and do so in efficient and appropriate language.

My interview data suggests that stance choices made by the writers in this corpus are not always appropriate. We do not know, of course, whether the issue is a lack of pragmatic awareness, or a lack of language resources with which to express desired pragmatic meanings. The textual data itself suggests that within each stance option, some language resources are particularly popular. However, these resources may, from a disciplinary perspective, not be completely appropriate. For example, my informant commented on the phrase 'on average life expectancy is about 75 years', noting that it was appropriate for the student to hedge the assertion, but that the phrase 'on average' was strange because life expectancy is always an average. A pedagogic goal, then, is to help students expand the repertoire of language resources with which to express stance options.

Both pragmatic goals and language repertoire goals could be addressed by engaging students in work on the same corpus as we have used for our research. Typically, perhaps, corpus based EAP work has involved teachers and/or students in looking at an aspirational corpus, something representative of the kinds and quality of language and/or texts that students are aiming to produce. (E.g. Chang & Kuo, 2011; Chang & Shleppegrell, 2011). I would agree that there is

much of value in such an approach, although examples of proficient student writing in Statistics are hard to find.

It has also been argued that learner corpora have a significant role to play (Aston, 2000; Flowerdew, 1998; Gilquin, Granger & Pacquot, 2007; Luzon, 2009), and I suggest that there is much that Statistics students could learn from discussing the writer choices represented in this small data set. The data lends itself to a strongly consciousness-raising approach, with students investigating language produced by their peers and being guided to evaluate its appropriacy. We can make use of our students' status as relative insiders in the micro, place-discourse community (Swales, 1998) that is the Stats Lab course at the University of Warwick. We can draw on their knowledge of writing in the discipline of Statistics, together with their knowledge of the expectations of their teachers, to help them to make appropriate and functional language choices and give themselves an effective voice. For example, two sets of factors influencing choices as to stance – social and epistemological – should be brought to students' attention. Then students, as community insiders, can be asked to consider both the epistemic and the pragmatic factors which might influence a writer to make certain linguistic choices.

My informant commented that many Statistics students dislike written tasks such as this one, because they think of Statistics as a mathematical subject and want to focus on learning more, and more difficult, ways of doing analysis. Lecturers, however, want students to go beyond the maths; they argue that it is equally important for students to be able to explain what they are doing in language. It seems to me that a learner corpus would be especially motivational in this

regard, with students able to investigate language produced by their peers in tasks directly comparable to those they will undertake. Further research is planned to investigate the effectiveness of such pedagogy, as well as examining other aspects of the learner corpus discussed here.

References

Aston, G. (2000). Corpora and language teaching. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora.* (pp. 7-17). Hamburg: Peter Lang.

BAWE Corpus Manual. Downloaded from Oxford Text Archive, http://ota.ahds.ac.uk/

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5,* 97-116.

Braine, G. (1989). Writing in science and technology: An analysis of assignments from ten undergraduate courses. *English for Specific Purposes, 8,* 3-15.

Bruce, I. (2009). Results section in sociology and organic chemistry articles: A genre analysis. *English for Specific Purposes, 28,* 105-124.

Chang, C-F. & Kuo, C-H. (2011). A corpus based approach to online materials development for writing research articles. *English for Specific Purposes, 30,* 222-234.

Chang, P. & Schleppegrell, M. (in press). Taking an effective authorial stance in academic writing: Making the linguistic resources explicit for L2 writers in the social sciences. *Journal of English for Academic Purposes* (2011), doi: 10.1016/j.jeap.2011.05.005

Channell, J. (1994). *Vague language.* Oxford: Oxford University Press.

Chen, H. (2010). Contrastive learner corpus analysis of epistemic modality and interlanguage pragmatic competence in L2 writing. *Arizona working papers in SLA and teaching, 17,* 27-51.

Eggins, S., (1994). *An introduction to systemic functional linguistics*. London: Continuum.

Flottum, K., Tinn, T., & Dahl, T. (2006). 'We now report on' versus 'Let us now see how': Author roles and interaction with readers in research articles. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines*. (pp. 203-224). Bern: Peter Lang.

Forster, M. , Smith, D., & Wild, C. (2005). Teaching students to write about Statistics. Paper to Statistics Education and the Communication of Statistics, Sydney, Australia, 4-5 April 2005. Downloaded from http://www.stat.auckland.ac.nz/~iase/publications/

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP

pedagogy. In P. Thompson (Ed.), *Corpus based EAP pedagogy. Special issue of Journal of English*

for Academic Purposes, 6, 4, 319-335.

Hsieh, H. & Shannon, F. 2005. Three approaches to qualitative content analysis. *Qualitative*

Health Research, 15 (9): 1277-1288

Hyland, K. (2005). *Metadiscourse*. London: Continuum.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. London:

Routledge.

Hyland, K. & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal*

of second language writing, 6, 2, 183-206.

Koutsantoni, D. (2004). Attitude, certainty and allusions to common knowledge in scientific

research articles. *Journal of English for Academic Purposes, 3,* 163-182.

Koutsantoni, D. (2006). Rhetorical strategies in engineering research articles and research

theses: Advanced academic literacy and relations of power. *Journal of English for Academic*

Purposes, 5, 19-36.

Lipson, K. & Kokonis, S. 2005. The implications of introducing report writing into an introductory statistics subject. Paper to Statistics Education and the Communication of Statistics, Sydney, Australia, 4-5 April 2005. Downloaded from http://www.stat.auckland.ac.nz/~iase/publications/

Luzon, M-J. (2009). The use of *we* in a learner corpus of reports written by EFL engineering students. *Journal of English for Academic Purposes, 8,* 192-206.

Lyons, J. (1977) *Semantics, Vol. 2.* Cambridge: Cambridge University Press.

MacGillivray, H. (2005). Helping students find their statistical voices. Paper to Statistics Education and the Communication of Statistics, Sydney, Australia, 4-5 April 2005. Downloaded from http://www.stat.auckland.ac.nz/~iase/publications/

Martin, J. & White, P. (2005). *The language of evaluation*. Basingstoke: Palgrave MacMillan.

McGrath, L. & Kuteeva, M. (2011). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. In press, *English for Specific Purposes.* doi:10.1016/j.esp.2011.11.002

Parke, C. (2008). Reasoning and communicating in the language of Statistics. *Journal of Statistics Education 16,1.* Downloaded from

www.amstat.org/publications/jse/v16n1/parke.html

Richards, K. (2003). *Qualitative inquiry in TESOL*. London: Palgrave.

Schönfelder, Walter (2011). CAQDAS and Qualitative Syllogism Logic—NVivo 8 and MAXQDA 10 Compared [91 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 21, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101218. Retrieved from

http://www.qualitative-research.net/index.php/fqs/article/view/1514/3134

Seror, J. (2005). Computers and qualitative data analysis: Paper, pens, and highlighters vs. screen, mouse and keyboard. *TESOL Quarterly, 39,2,* 321-328.

Swales, J. (1990). *Genre analysis.* Cambridge: Cambridge University Press.

Swales, J. (1998). *Other floors, other voices: A textography of a small university building*. Mahwah, NJ: Lawrence Erlbaum.

Vellenga, H. (2004). Learning pragmatics from ESL & EFL textbooks: How likely? *The Electronic Journal for English as a Second Language, 8, 2,* retrieved from http://tesl-ej.org/ej30/a3.html.

Wu, S.M. (2007). The use of engagement resources in high- and low-rated undergraduate

geography essays. *Journal of English for Academic Purposes, 6,* 254-271.

Zhang, Y. & Wildermuth, B.M. (2009). Qualitative analysis of content. In B. Wildermuth (Ed.),

Applications of social research methods to qualitative studies in information and library.

Retrieved from http://www.ils.unc.edu/~yanz/Content_analysis.pdf.