Single language corpus, multilingual background

Geoffrey Williams, Claude Sionis and Paul Boucher

Electronic version
URL: http://asp.revues.org/1479
DOI: 10.4000/asp.1479
ISSN: 2108-6354

Printed version
Date of publication: 30 December 2002
Number of pages: 47-57
ISSN: 1246-8185

Electronic reference

This text was automatically generated on 21 February 2017.

Tous droits réservés
Single language corpus, multilingual background

Geoffrey Williams, Claude Sionis and Paul Boucher

1 Publishing is a major headache for the non-native speaker of English (NNSE) researcher. It is essential both for the propagation of scientific knowledge and for career advancement as a scientist is judged by the number of articles published, and the prestige of the journal. For most sciences there is no alternative to publishing in English language journals as even if other outlets exist, their rating will be low. This need to write and publish in English is where the stumbling block of language competence arises. Many NNSE feel that they are being excluded from the world scientific community solely on linguistic criteria.

2 To get published the non-native writer may adopt one of three strategies: translate, or have translated, the paper, get it checked by a native speaker, or improve his or her own command of English so as to produce acceptable English text. The first is obviously impractical, and far too expensive. The second means being reliant on a willing third party, which leads the third as the preferred choice.

3 English for Specific Purposes/English for Academic Purposes (ESP/EAP) teaching seeks to teach the skills necessary for the NNSE writer. Work in ESP/EAP has meant that a considerable body of literature has been built up over the years based on the analysis of real texts. However, these studies have generally used fairly small collections of texts, which means that it is often impossible to generalise their conclusions (see Busch-Lauer in this volume). The rise of corpus linguistics now means that large amounts of data can be studied using computers; a corpus is not measured in tens of texts, but in hundreds. However, the use of specialised corpora in language research can bring new problems of over-generalisation, which can mean that what becomes accepted as a norm may only be the effect of avoidance strategies.

4 This paper does not seek to answer precise questions as to NSE and NNSE usage, but rather to outline a data-driven approach to corpus analysis of genre-specific discourse. It starts out by demonstrating the danger of hasty judgements as to NSE and NNSE status in
corpus studies and goes on to consider usage of one factor, the which/that choice in relative clauses.

**NNSE and the specialised corpus**

Corpora are large bodies of texts assembled according to explicit linguistic criteria (Sinclair 1996). For lexicographic purposes very large reference corpora such as the Bank of English or the British National Corpus have been assembled. These aim to be representative of the language at a given moment in time and are carefully balanced as to context (Atkins et al. 1992). In such corpora any individual variations from the grammatical norm are lost due to the sheer size of the corpora. However, the problem with reference corpora is that they give a general overview of the language and cannot be relied on for genre-specific studies. At the other end of the scale, small specialised corpora of about one million words in size can guarantee both domain and genre specificity, but bring in other problems of representativity (Williams 2002). In the assembling of specialised corpora, representativity is seen in terms of field or topic specificity, but this in turn leads to another important aspect being overlooked, that of the first language status of the contributors (Williams 2002).

A recent approach to the study of NNSE writing has been the development of learner corpora (Granger 1998). In building learner corpora, a series of comparable corpora are constructed with writers from known language backgrounds writing on set subjects. The resulting corpora can then be compared to a benchmark corpus built from the production of mother tongue writers working under the same conditions as the NNSE writers. The results are extremely interesting and may transform teaching manuals; however, the technique is not applicable to specialised corpora where real research papers are required to study a living genre. One solution might be to add information as to language background in the corpus header, which would make sub-categorisation of the corpora possible. However, such a procedure would introduce considerable bias, and begs the question as to who really is a native speaker.

If we were to limit a specialised corpus to only those writers “known” to be living and working in the UK, the results would be so skewed as to be useless. Apart from having lost the topic specific information we are seeking, a surname cannot tell us whether an author was born, or educated, in an English speaking country. In multicultural societies, such as in Britain or the United States, surnames tell us little. We cannot, for instance, judge from a name whether someone is a second or third generation immigrant who speaks no language other than English. This is further complicated by varieties of English as on the Indian subcontinent and large parts of Africa and Asia English is the lingua franca. We can no more write off these varieties as non-standard than those of Australian or American English. Another problem in subcategorising a corpus would be the tendency to have multiple authors for a scientific research paper. It is impossible to say which author really wrote the article and the role played by the other authors. Another question is how to deal with, for instance, a German or French post doctoral student working in a US laboratory. Do we class the paper as a native or non-native production? The possibilities are numerous, and further compounded by the language competence of the native speaker. Many NNSE writers will have had a far more rigorous instruction in grammatical usage than their English counterpart.
Clearly the only solution is to accept a corpus as a whole and rely on the fact that peer reviewed journals will have forced contributors into bringing their paper up to a publishable standard. However, this too is problematic as while plain bad usage and sloppy punctuation may disappear, there will also be use of avoidance strategies leading to correct, but rhetorically poor, texts.

Having isolated some potential problems of corpus analysis, we can now look for solutions by developing research paradigms that take into account the shortcomings of corpora whilst putting their strengths to the service of the NNSE research writer.

Overcoming avoidance strategies

Dealing with basic grammatical infelicities is a relatively simple operation; more complex is the problem of reformulation. Most writers are aware of their shortcomings and attempt to overcome potential problems by the use of avoidance strategies. This can mean, for example, attempting to avoid error by recourse to simple forms, as juxtaposition. Such avoidance leads not only to rhetorically poor texts, but frustration for the writer, who is unsatisfied with the text stylistically, and often afraid that the scientific content may be affected.

The answer is not prescriptive formulae, which often lead to heavy, inexpressive texts, but to isolate the potential factors of poverty from genre-imposed restrictions and to present possible reformulations. Once the task of analysis has been achieved, it might be possible to develop a tool to supply some form of automatic assistance. Consequently, the aim here is not to criticise papers, but to attempt to use a corpus to isolate factors of poverty so as to develop tools with which to assist in the writing process.

This is to be done by applying both top down and bottom up approaches within a transdisciplinary research group that will bring together the insights of both theoretical and applied linguistics. The three main approaches are those of the individual researchers, looking at the same material and attempting to produce both a synthesis and practical results.

The Analyse Linguistique et Pratiques Langagière research laboratory is led by three linguists involved in ESP/EAP research: Claude Sionis (Université de La Rochelle), who specialises in the pragmatic aspects of research writing, Paul Boucher (Université de Nantes), primarily concerned with the syntactic analyses of written language and Geoffrey Williams (Université de Bretagne-Sud, Lorient), who researches in corpus-driven lexicography. The team also works closely with other researchers in the field of scientific writing and specialists, particularly in Natural Language Processing (NLP), concerned with multimedia, distance learning and terminology extraction. What gives the team coherence is the fact that the three leaders are, in addition to their research activities, involved in teaching writing strategies to both students and experienced scientists. The other factor is that we are dedicated to using corpora as a means of studying real language in context.

Looking at language

One way of approaching usage is through what has been termed the corpus-based approach in which a corpus is used to check out previously formed hypotheses. If we start
with the hypothesis that NNSE writing may be ill-formed or syntactically poor due to avoidance strategies, we would then classify the files by origin of author. As we have seen, two problems arise with this approach: we return to the problem of defining a native speaker, and we face the standard problem of all hypothesis based approaches to corpora: if you set out to look, you will almost certainly find, but pass by all the interesting details on the way. Our preference is for a corpus-driven approach (Tognini-Bonelli 2001) in which the data is allowed to speak for itself, even if this does remove that sense of security supplied by the simple need to confirm a pre-established hypothesis.

In a corpus-directed perspective, rather than attempting to isolate what is native and non-native speaker writing, we start with a quantitative analysis so as to group texts without reference to authors’ linguistic origins, but only as a function of linguistic strategy. This has to be handled with care as we all know the dangers of factor analysis in corpora. It is not sufficient to know that something is present; we need to know what purpose, rhetorical or syntactic, it is serving. To do this, both top down and bottom up analyses are used to offset corpus findings against those of a pragmatician and a syntactician, and vice versa. The texts are analysed using corpus tools to discover regularities, the results are then analysed by a syntactician with reference to accepted norms and by a pragmatician to study for what reason the norms may have been adapted or flouted.

This is a deconstructive stage that must be carried out before reconstructive means to assist with research writing can be envisaged. The aim is to always avoid the simplistic prescriptive methods found in self-help books by demonstrating a variety of strategies in context. Given the width of the field we deal with, we initially intend to tackle only one feature, the use of the relative pronouns which and that. The aim is not to make a definitive statement on the which/that choice, but to illustrate a data driven approach to this question.

The corpus

In order to carry out this project a variety of sources are being built, notably a learner corpus of pre-publication articles in English coming from Chinese, French, German and Russian speakers. In the work described here the main source is the BIVEG corpus (Williams 2001). This resource was originally built to look at the interaction between language use in plant physiology and plant molecular biology and, up to now, has primarily been used for lexicographical purposes. Rather than taking the corpus as a whole, we are looking at two subsets that cut across disciplinary boundaries: the proceedings of a conference on parasitic plant biology held in Cordoba in 1996 and the articles from peer-reviewed journals.

The reason for this breakdown is simple. As anyone working with NNSE researchers knows, editors tend to nitpick on grammatical details, so we could presume that peer-reviewed articles will be relatively well-formed. On the other hand, the peer-review process for conferences is often essentially scientific with much of the selection carried out by the organising committee, NNSE or not. This was indeed the case for the Cordoba conference, and in this instance the committee most certainly did not want their own texts looked at for grammatical idiosyncrasies.
Lexicographical research on BIVEG has already shown that there is a tremendous difference in terminological density between the two subsets in that one is addressing a field specific community through a peer reviewed journal, the other a topic based community which is multidisciplinary. The question which we have set out to look into here is whether there are more syntactic infelicities in the topic based subset that are reflected in syntactic poverty in the other group when NNSE writers use avoidance strategies to get round difficulties of expression.

The grammatical preferences to be found in scientific genres are well known thanks to the active research in English for Academic Purposes from the pioneer work of Barber (1962) to the present day. The problems facing NNSE writers are equally known. Structural problems are often pragmatic and expressed through the organisation of information, essentially theme–rheme structure. Grammatical problems include the use of relatives, determiners, compounding and tense choice. These too are largely pragmatic in nature as they relate to choices within a restricted socio-rhetorical context. Writers often attempt to get round perceived difficulties by recourse to avoidance strategies. We know, for instance, that it is easier to juxtapose than to create complex structures so relatives might be seen not only as important in information structure but also as a step towards compounding.

The which/that choice

As mentioned, factor analysis is not a sufficient criterion on its own to typify genre-specific grammatical choices as it begs the question as to why an author is using a particular structure. Simple quantitative analyses fail to tackle pragmatic issues and would thus render a study such as this of little value. So, rather than just counting, we propose to begin with that and which as markers of relativity and to then group the texts by order of frequency of these markers in order to isolate high, low and mid frequency users. These subsets would then be studied to see how and to what purpose these relatives are being used and what alternative strategies are being adopted. The subsets extracted will be small, only five texts in each category, but this is sufficient for a pilot study.

In such an inductive approach we can then go back to the initial corpus subset to look at the frequency of alternative rhetorical and syntactic strategies so as to see to what extent avoidance strategies are being used. Further work will be required to see whether it is the high frequency groups that are using avoidance strategies to avoid, for example, compounding.

“Which”

“Which” (graph one) has a wide range, going from a high usage to three texts which do not use it at all. A similar range can be seen for “that” (graph two). From this it might be surmised that some authors have a preference for one or the other, which would mean that these pronouns are relatively interchangeable. This is not, of course, the case, firstly, and most obviously, because of constraints on restrictive and non-restrictive clauses.
If we take both the high and low scoring groups, the five texts with the highest and lowest frequencies, we find an equal use of restrictive and non restrictive clauses. In the low frequency group, it is rare to find “that” incorrectly used in non restrictive clauses, but incorrect use of punctuation is frequent, as can be seen in table 1.

Table 1. Anomalies in the low frequency group

The parent line N-13 which is known to possess moderate resistance to Striga under field conditions, had high-stimulant characteristics. The infestation totals less than 3000 m2 which we fumigated with methyl bromide.

Nuclei of Striga were predominantly euchromatic in contrast to those of Vicia which appeared heterochromatic.

There are also differences in usage. If we take this low frequency group we find that low frequency does not automatically equate with NNS usage. It does, however, equate with
texts that are largely descriptive and factual, texts that are simply reporting rather than discussing the implications of the findings (table 2). To carry out this part of the analysis we have ‘cheated’ by looking at the corpus headers to see whether any known native speaker single author papers are present in the group. This was the case for two articles. We can also note that although all the texts use complex sentence structures, two of them have poor cohesion and make use of complex juxtaposition. Thus in this case we seem to be witnessing a lack of rhetorical strategies coupled with problems of usage. The grammatical strategies do not seem to be weak as these texts displayed a wide use of non-finite clauses.

Table 2. Rhetorical and grammatical structure in the low frequency group

<table>
<thead>
<tr>
<th>Text Code</th>
<th>Rhetorical and Grammatical Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BV054PPA</td>
<td>Descriptive, factual, complex juxtaposition, poor cohesion</td>
</tr>
<tr>
<td>BV067PPV</td>
<td>Descriptive, factual, complex juxtaposition</td>
</tr>
<tr>
<td>BV074PPA</td>
<td>Descriptive, factual, complex structures, cohesive - NSE</td>
</tr>
<tr>
<td>BV090PPP</td>
<td>Descriptive, factual, complex structures, cohesive - NSE</td>
</tr>
<tr>
<td>BV101PPV</td>
<td>Descriptive, factual, complex juxtaposition, poor cohesion</td>
</tr>
</tbody>
</table>

If we take the high group (table 3), we find a different picture. Here the texts are no longer simple description. There is also on the whole a much better use of cohesive devices. It would be hasty to draw conclusions from such a small sample, but two facts seem to be coming clear: use of relatives is linked to text purpose, purely factual to the descriptive and that the purely factual may in some cases be the result of avoidance strategies. To check this out we must turn to the peer reviewed subset, and here very similar patterns are to be found. In this peer reviewed subcorpus, we also found a surprisingly high misuse of non-restrictives, which must show that reviewers are not always as draconian as imagined, and that attested scientific usage may not be purely the grammatical specificities of so-called sublanguages, but largely the result of limited grammatical and structural strategies.

Table 3. Rhetorical and grammatical structure in the high frequency group

<table>
<thead>
<tr>
<th>Text Code</th>
<th>Rhetorical and Grammatical Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BV053PPP</td>
<td>Classifying by comparison, high use of adverbials, cohesive - NSE</td>
</tr>
<tr>
<td>BV059PPV</td>
<td>Descriptive discussion, poor punctuation of non-restrictives</td>
</tr>
<tr>
<td>BV060PPP</td>
<td>Descriptive discussion, high use of adverbials, cohesive - NSE</td>
</tr>
<tr>
<td>BV069PPV</td>
<td>Classifying by comparison, well structured, cohesive</td>
</tr>
<tr>
<td>BV096PPV</td>
<td>Experimental, well structured, cohesive - NSE</td>
</tr>
</tbody>
</table>

Problem Solving

Having looked at what the corpus reveals on which/that usage, what we are beginning to find is that we must differentiate two problems, ill-formed phrasing and avoidance strategies. The former could be dealt with through some form of grammar checker that has been calibrated to take into account usage in the field of the writer. The latter is more complex in that a grammar corrector only looks at what is “wrong”, not what is missing. So, linked to the grammar checker we should have a tool that will offer reformulations. Such a tool would have to be rendered aware of specificities of sub-genre in research writing. It would also need to be taught new functions, possibly in the same way as a translation memory learns. This is something that we will be considering in the longer term.
However, before we can teach we must clarify our theoretical position. If one group prefers “that” to “which”, or vice versa we have to know whether there is a difference in usage within restrictive clauses.

One possibility is that the difference is pragmatic: the anaphoric/cataphoric opposition. This can be explained theoretically, but in the corpus examples there is little evidence of it in practice. In many cases the choice appears completely open, within restricted clauses, that is. The fact that a word processor such as Microsoft Word prefers “that” for “which” in its grammar corrector is hardly a reference, the preference amongst British or British-educated sources is very much for “which”. It is possible that there is little theoretical reason. We are simply in the domain of colligation, restricted uses.

In the monumental *Comprehensive Grammar of the English Language*, Quirk et al. (1985) report no difference of use in general restricted clauses, but a number of differences in certain precise cases (17.15: 1250). It is then necessary to see whether these restrictions are respected or whether the form is avoided.

Norm 1

“That” or zero is preferred for non-personal antecedents, all, everything...

There are only two examples of the former. The first is a mistake in an NNSE text where much less is followed by ‘that’ rather than ‘than’ and:

Anything which can reveal allelic ...

The latter is an NSE production, which could mean that either this formula is not scientific usage, or that we cannot always trust the NSE writer.

Norm 2

“That” is preferred when the antecedent is modified by a superlative or by the post determiners first, last, next, only

This second formula revealed the phrase “in the first true leaves which were at an advanced” from a mixed lab but with one NSE author. As regards superlatives only one was found, and that did conform to the rule.

figure 1 represents the most parsimonious tree that was constructed ...

In this case we cannot tell whether the lack of data corresponds to an avoidance strategy or simply that these formulae have only a limited place in scientific writing. Only reference to a larger corpus will tell, and if it is an avoidance strategy then means must be found to exemplify usage so as to widen the rhetorical arsenal of the NNSE writer.

The obvious difference in usage of “which”/“that” is in relative and appositive clauses. The latter do not seem to be a problem as there is no element of choice; these are lexical phrases in their own right of the form the fact that, there is evidence that etc. This concerns a relatively closed set as “indication”, “observation(s)”, “speculation”, “suggestion”. In noting patterns we must be aware of traps as in the observation that and an observation which/that. This brings us back to the related problem of determiners, and in particular the definite article. As Quirk et al. point out there can be an ambiguity between relative and appositive ‘that’, but it is rare. Misuse of an appositive would be
immediately noticed and corrected by reviewers, but the NNSE writer would rather know before.

37 We have reached the point where we can say that that and which relative pronouns are interchangeable, except where they are not. Somehow the restrictions, which are usage restrictions rather than pragmatic or syntactic, will have to be exemplified. Relatives in this corpus are mostly used to refine rather than define an idea. If the anaphora/cataphorical is real and can explain preferences for one or other form, then we are going to have to find how this is expressed as there is no point in explaining this to the user.

38 This all goes to prove that you cannot teach by theory or rule, but only by example, and to do that we come to the dictionary as teaching tool and the representation of grammar within the specialised lexical context.

39 There are two ways of tackling the problem of introducing new formulae: one is through a grammar section, the other is to link the formulae to the lexis. We envisage using both, but by emphasising grammatical patterns (Hunston & Francis 1999) which supply the contextual meaning of lexical items. Among the grammatical patterns that emerge in this corpus are those related to what Quirk et al call relative as adverbial, that is the preposition + which forms. It does appear that certain authors avoid these forms.

40 In the high/low groups for which/that relatives, the use of adverbials fell into the higher groups and the most frequent usage was exclusively among NSE writers. Faced with the question as how to exemplify these, the answer seems to be through the head noun or its hypernym.

41 Prepositions in relative clauses are fulfilling precise functions relating to the means whereby something happens as in “by”, relating to time or duration as with “after” and “during”, etc. Within this corpus these functions are often related to a limited lexis. For instance “by” occurs with mechanisms and processes, so the clause could be introduced through the hyponym “process”. This is defendable in that although the “by which” pattern is not frequent in comparison with the occurrences of “process” and its hypernyms, the co-occurrence pattern is significant if we work back from “by which”. In the corpus the following patterns appeared as being significant see table 4).

<table>
<thead>
<tr>
<th>Table 4. Significant patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>• At which</td>
</tr>
<tr>
<td>- “frequency” at which x occurs</td>
</tr>
<tr>
<td>- “temperature” at which x occurs</td>
</tr>
<tr>
<td>- “place” at which x occurs</td>
</tr>
<tr>
<td>• By which</td>
</tr>
<tr>
<td>- “process” by which X happens</td>
</tr>
<tr>
<td>• Of which</td>
</tr>
<tr>
<td>- “number” of which</td>
</tr>
</tbody>
</table>

42 Such formulae would then have to be demonstrated along with the other grammar patterns found with the head word, and the headwords linked through their classifier to demonstrate the regularity of the pattern.

43 This is only a passive offer. It goes without saying that a more active reformulation process would offer greater assistance to the apprentice science writer. Further corpus-driven analyses will be necessary to locate potential problem areas.
Determiners have only been mentioned in passing as these represent a very different problem, a matter of pragmatic choice. Definite articles do pose a particular problem for French speakers of English. When correcting texts we tend to add missing articles, however, if the text is reread by a colleague, then he or she will eliminate some of what has been added. Again the only teaching answer will be by exemplification, but a rigorous pragmatic analysis will also be required in that the decisions are not syntactically based.

Conclusion

At a recent meeting of linguists in Marburg, the chairman announced that “The English of Science is bad English”. This is inevitable in the development of World Englishes where the aim is above all to communicate. This is, after all, the main purpose of language. However this does raise questions with benchmarks as to publishable standards in science writing.

If we draw up a list of grammatical functions that “typify” science writing, are we reflecting a developing socio-rhetorical strategy adopted for reasons of precision and scientific rigour, or simply a series of avoidance strategies that simply reflect the needs of NNS scientists to get their word in edgeways? The answer may be to look at corpora to identify avoidance strategies and seek ways to alleviate them, within the norms of science, not trying to impose the model of literary practice.

Once patterns have been isolated on small corpora it will be necessary to see whether generalisations can be drawn. This means looking at other corpora, which means building more specialised corpora. The answer is not reference corpora which are too expensive and too heterogeneous. Existing specialised corpora cannot be enlarged without the risk of losing that very topic centredness that is essential in understanding languages for specific purposes. This is by necessity ongoing research. Language does not stay still, so writing strategies will have to be adapted to current trends. No solution can ever be seen as definitive. What we are putting forward here is an approach to specialised corpora that harnesses the power of computer technology without losing sight of the human aspects of text production. A totally automatic approach can only lead to standardised solutions, which in turn reintroduce a form of poverty that we are seeking to overcome. A corpus-driven solution is not an easy one as it forces the linguist to remain close to the uncomfortable reality where exceptions are the norm. On the other hand, writers, whether of scientific papers or other genres, do not seek to produce mechanical texts, they seek to be read. The grammatical exceptions are the inconveniences that render a text of interest.

BIBLIOGRAPHY

ABSTRACTS

Publication is a major problem for the non-native speakers. Getting texts checked by a native speaker of English is not always feasible, so other strategies must be found. The study demonstrates that in corpus composition it is impossible to isolate native speaker production, the sheer mass of data, at least 500,000 running words, in a correctly assembled corpus renders this unnecessary. This text calls for a data driven analysis of scientific corpora so as to isolate the norms acceptable to a language and discourse community. This is shown with a study of which/that clauses. It concludes that perceived norms may be influenced by avoidance strategies.

La publication reste un problème majeur pour les rédacteurs non-natifs. Faire contrôler des textes par un locuteur natif n’est pas toujours possible, d’autres stratégies doivent donc être trouvées. Cette étude démontre que, dans la composition d’un corpus, il est impossible d’identifier les productions des locuteurs natifs, mais que la masse de données dans un corpus scientifique correctement constitué, un minimum de 500 000 mots, rendra cette opération inutile. L’étude propose une analyse inductive d’un corpus scientifique afin de dégager les normes admises par une communauté de langue et de discours. L’approche repose sur une étude des formes which/that qui démontre que les normes perçues peuvent être influencées par des stratégies d’évitement.

INDEX

Keywords: corpus, data-driven analysis, non-native speaker, scientific
Mots-clés: analyse inductive de corpus, corpus, rédacteur non natif, scientifique
AUTHORS

GEOFFREY WILLIAMS
CLAUDE SIONIS
PAUL BOUCHER