Mat Plews

Measuring the effectiveness of a task-based spoken EAP instructional sequence

1. Introduction

Within the field of English for Academic Purposes (EAP), task-based language teaching (TBLT) is commonly considered best-practice, with broad claims being made that it promotes learning and replicates the academic discourse community (Gillet/Wray 2006: 6-7, Richards/Rogers 2001, Feez 1998, Charles 1996), promotes critical-thinking (Alexander et al 2008: 267), and is particularly suitable for high-level learners (Williams 2001, Long/Porter, 1985). However, there are also those who question its general utility (e.g. Bruton 2002, 2010; Swan 2005), arguing that it does not promote language development as much as its supporters claim.

With regard to actually measuring the effectiveness of a TBLT approach in an EAP context, however, there has, thus far, only been limited, almost exclusively qualitative, research related largely to the effectiveness of entire, task-based EAP courses (Gillet/Wray, eds 2006; Storch/Tapper 2009), using questionnaires, interviews or pre- and post-course tests. However, there has as yet been no study which has attempted to link classroom activity directly to outcomes in a quantifiable manner, clearly desirable if it is to be accurately determined whether and how learning actually takes place. This is precisely what this research project intends to do, in order to contribute to the debate on whether TBLT deserves its 'best-practice' label in the EAP context.

The focus of the analysis in this study is a task-based sequence of spoken classroom debates on the topic of the 'Israeli-Palestinian conflict', conducted by a group of mixed-nationality B2+-level (see Council of Europe, 2008) university students as part of an EAP course in 'English for Politics and Economics' at the Humboldt University Language Centre, Berlin. The aforementioned sequence of debates conforms to Ellis' (2009: 223) requirements for 'tasks' in that each discussion as well as the whole has a "… clearly-defined outcome other than the use of language …" (in this case, the objective of agreeing on proposals aiming to solve the conflict), the students "… largely have to rely on their own resources …", and the "… primary

focus is on 'meaning' …" In more detail, this 'instructional sequence' (a term borrowed from Lynch 1996) involved an initial student debate (T1), followed by a 'focus-on-form' phase (FoF, in which students compared their performance in T1 with that of native speakers conducting a debate on the same topic) and a second, similar, student debate (T2). All three phases were digitally recorded for analysis. Structuring the sequence in such a way meant that students performed tasks and received input at the i+1 level (Krashen 1982), with possibilities to identify the language that they needed themselves (Schmidt 1990) and to try it out i.e. 'negotiate meaning' (Long 1983) in the final debate, thereby testing the key theories underpinning TBLT.

In order to quantitatively determine the effectiveness of this task-based spoken EAP instructional sequence, lexical-analysis software was used to measure the number and length of all new 'linguistic items' (words and lexico-grammatical chunks) from the FoF which were reused by students in T2, thereby directly connecting any linguistic gains to classroom activity.

2. Methods

For the debates, the class was divided into two groups, one to put forward the Palestinian viewpoints and the other the Israeli arguments. In order to obtain a balanced mix of individual variables for analysis, the groups were divided as equally as possible in terms of C-(entrance) test score, nationality, gender, and my own view of their motivation.

With the aim of providing 'interactional authenticity' (Bachman 1990) to promote natural use of language appropriate to an academic context, prior to the initial debate, the students were asked to complete background reading and viewing assignments on the Israeli-Palestinian conflict related to ten key areas and, in their groups, to share ideas with regard to these.

In the first debate (T1), lasting just under one hour, the ten areas were discussed in turn, with all students being allowed to contribute at any time. My role was limited to chairing the discussion although, on occasion, I 'recast' what I judged to be important language mistakes by providing the correct version.

Following the initial debate, for homework, was the 'focus-on-form' stage, in which the students listened to a recording and looked at the transcript of native speakers performing the same task, as many times as they liked. While doing this, they were asked to note down / highlight what they considered to be the positive and negative points concerning the way in

which the native speakers expressed themselves in comparison to the students' own performances. Audio-files of both the native-speaker and T1 recordings, as well as the transcript, were available for home-use by students, on the course's Virtual Learning Environment (Moodle). This activity led to students annotating and reorganising the information on their copy of the transcript to produce a document containing 'useful language' for the next debate (T2), which they were allowed to use for preparation, but not actually during T2.

One week later, the second student debate (T2, also recorded and just under one hour's duration) took place, with students swapping 'sides' and now putting forward the viewpoints of the other party, in order to consider the issues from another perspective, thereby providing further motivation for this phase (and ideas for a later essay on the topic). Before T2 began, students were asked for their opinions on the main differences between their first debate and that of the native speakers, and the teacher gave a brief overview of what he deemed these to be (without drawing attention to specific language). Students were also told, just before the second debate started, that they would not be allowed to use their 'useful language' documents, to ensure that any re-use of language from FoF came from experience and memory only, and to aim to prevent deliberate memorisation just before the debate started. The second debate, T2, also prefaced by the sharing of ideas within each group on the ten key areas, then proceeded in exactly the same way as T1.

Following this final debate, in order to determine whether students had re-used any of the native speakers' language, the transcripts of T2 and the native speakers' debate were compared and any lexical 'items' (words or chunks) occurring in both were noted. If any of these items also occurred in T1, then they were removed, leaving a list of lexical items occurring in T2 and the FoF only. This "new" language was then compared with all of the language in the FoF but not in T1, to determine to what degree native-speaker language had been reused and could thus be judged to have been acquired (at least in the short-term). Before the whole analysis began, the words spoken by the Chair in both student debates (i.e. the teacher) were removed. The Chair's words from T1 were then added on to the transcript of the native speakers, as this was also language that students could have acquired from a native speaker in the first debate, and therefore part of the focus-on-form.

Alongside these results for the group as a whole, it was deemed necessary and potentially instructive to also focus on individual results, as there is clearly a large amount of individual variation amongst students (Skehan

2009: 523, Bruton 2010). For this reason, individual results and case-studies were also produced.

To process and analyse all the data, lexical analysis tools from Tom Cobb's *The Compleat Lexical Tutor* (http://www.lextutor.ca/) were used, alongside the 'Find' functions in *Microsoft Word* and *Excel*.

In detail, *Text-based Range* was used to detect the reoccurrence of individual words and word families, producing lists in an *Excel* spreadsheet for all items in the three debates. From these lists, those items which were judged to be significant were those occurring at least twice in the native-speaker debate, and details of those occurring in both FoF and T2 only were obtained simply by removing all those occurring in T1.

The tool *N-Gram Phrase Extractor* was useful for producing lists of all significant (i.e. occurring twice or more) 2- to 5-word strings for each debate. Those for 2-, 4-, and 5-word strings in the native-speaker debate were used as search items in *Word* to identify those occurring in T2 but not in T1. In this way, lists of all the new significant strings occurring in T2 were identified.

For 3-word strings, the procedure was simpler. Instead of having to search using *Word*, *Text Lex Compare* (which, unfortunately, does not produce lists for 2-, 4-, and 5-word strings) was used to detect significant strings in FoF which also occur in T2. These were then compared with T1 to produce a list of those occurring in both FoF and T2 only.

The numbers of individual words, family words ('root' words for a family e.g. occurrences of "*conceded*", "*concession*" and "*concedes*" only appear once under the family word "*concede*", although they appear as three separate individual words) and 2- to 5-word strings in each final list (i.e. the significant words from FoF occurring in T2 only) were then expressed as percentages of the respective total number of 'new' significant items in FoF (i.e. all those not present in T1) to provide final percentages for significant words and short phrases reused by students.

To detect strings of more than five words plus other 4- and 5-word strings occurring only once in FoF (any phrase of four words or more was judged to be significant), *Text Lex Compare* was used to produce a list of *all* 3-word strings occurring in both FoF and T2 only. *Word* was then used to compare their occurrences in these two debates, to see if they happened to occur within longer, more significant strings. It was not deemed useful to express the number of such strings occurring as a percentage of the total number of 'new' strings in FoF, as this latter number would be huge because the software counts every *n* words together as a string.

Following this process, searches were conducted in *Word* to produce results for each individual student.

3. Results

This section, concerned primarily with the results for the whole group, provides percentages of significant words and word-strings from the native-speaker (FoF) debate reoccurring in the second student debate (T2), as well as totals for the numbers of 'new' (i.e. not occurring in T1) 4- to 11-word strings reused in T2. With regard to individual results, the main findings are summarised at the end of this section.

Table 1 (below) shows the re-use of 'new' 'significant' words and word-strings from the native-speaker (FoF) debate in the second student debate (T2), expressed in percentage terms (for the purposes of this analysis, 'significant' items are defined as those occurring more than once in the FoF). Only exact matches are counted, except in cases where the language is so similar that almost no difference exists.

	New items recurring in FoF	Reused in T2	*Percentage re-use*
Family words	158	54	*34.18%*
Individual words (Types and Tokens)	234	73	*31.20%*
2-word strings	542	133	*24.54%*
3-word strings	391	65	*16.62%*
4-word strings	153	24	*15.69%*
5- word strings	53	9	*16.98%*

Table 1: Percentage re-use of new words and 2- to 5-word strings recurring in FoF

The results for reoccurrence of new 'significant' items clearly show some variation, with the analysis for Family and Individual words yielding figures of over 30% and those for 3-word-plus strings approximately the half of this.

To investigate this discrepancy, I decided to compare the total numbers of 'significant' items occurring in both student debates, T1 and T2. It was assumed that if these totals were reasonably similar, then this would render the corresponding percentage figure from Table 1 relatively meaningless, as it would show that such figures for re-use of lexis are normal even without a FoF stage. On the other hand, if the figures showed a remarkable rise from T1 to T2, then the corresponding row from Table 1 could be justified as important. The results of this investigation are shown in Table 2 (next page).

	T1	T2	*Percentage change*
Family words	310	323	*+4.19%*
Individual words (Types and Tokens)	357	360	*+0.84%*
2-word strings	392	424	*+8.16%*
3-word strings	109	133	*+22.02%*
4-word strings	11	28	*+154.54%*
5- word strings	3	11	*+366.67%*

Table 2: Percentage change for occurrence of 'significant' items from FoF in the student debates

It is clear from the investigation that the totals for Family and Individual words show very little variance, demonstrating that the relatively high figures for re-use of new single words from the FoF shown in Table 1 are insignificant. This is very probably due to the fact that in hour-long debates of the kind used in this study, so many words are used that there are bound to be a large number of common words. This also seems to be the case for strings of three words or less.

Conversely, the percentage changes for both 4- and 5-word strings are notably high, showing that the figures for re-use of longer strings in Table 1 are particularly important.

As results for lexical strings of four or more words (from now on also called 'longer strings') seem to be vastly more indicative of vocabulary acquisition than the shorter items, I decided to focus solely on these.

Consequently, the rest of this section is concerned only with figures for re-use of new 4+-word lexical strings in the second student debate, T2. Instead, however, of focusing only on items recurring in the FoF (up until now, termed 'significant' items), it is now clear that the items which are truly significant are all of those which are made up of four words or more.

The table on the next page (Table 3) presents totals for new 'distinct' (i.e. in the longest form in which they occur for an individual) 4+-word lexical strings also detected in the second student debate, including those which are recurring in the FoF and those which only occur once. Any strings recurring in the FoF which are only part of other, longer strings spoken by the same student (e.g. *"to take the first"* and *"take the first step",* which appear as two separate 4-word strings in the results in the section above) are only counted as one longer string (e.g. the 5-word string, *"to take the first step"*, appearing in both FoF and T2).

As the total number of recurring 4+-word strings from FoF also appearing in both T2 and T1 was relatively small (28 four-word and 11 five-word strings), and the single-occurrence (in FoF) strings were being searched for individually anyway, it made sense to concurrently note the names of the students who produced them, for later individual analysis. This also meant that new occurrences for individual students could be counted as new occurrences overall, even if one (or more) of their classmates had used the same string in the first debate e.g. although *"of the State of Israel"* was used twice by his classmates in T1, one student, PB (see Appendix) uses it for the first time in T2, inferring that he either acquired it from them or from the native speakers. As the focus-on-form stage involved students in analysing the class' performance in T1 as well as that of the native speakers, it is thus deemed justifiable to add reoccurrences such as this to the totals for new 4+-word string re-use. Strings which were reused by more than one student in T2, however, are only counted once in the whole class totals.

String-length	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Reoccurrences	41	16	3	2	1	1	2	1

Table 3: Number of reoccurrences of new distinct 4+-word strings from FoF in T2

A total of 67 new distinct strings of four words or more (listed in the Appendix) were detected as being reused in T2, which indicates that there was a significant amount of re-use of new phrases from the FoF by the students, with the vast majority (57) being 4- or 5-word strings, and there being only one occurrence each for 8-, 9-, and 11-word strings.

These results, coupled with the earlier figures of 16% and 17% for re-use of 4- and 5-word strings respectively appear to indicate that a significant amount of retention has taken place.

With regard to figures for re-use by individual students, a large degree of variation was found in the number and size of linguistic chunks retained. Hence, basic linear correlations were calculated between individuals' numbers of chunks retained and their C-test scores and numbers of words spoken in T2, producing results of r=0.7633 and r=0.8222 respectively, demonstrating strong degrees of correlation (a score of r=1 being a perfect correlation and 0 no correlation). The result concerning C-test score concurs with those of other TBLT researchers (Bruton 2010, Williams 2001, Long/Porter 1985), indicating that this approach particularly suits high-level learners, and the figure related to number of words spoken in the second debate implies that less-proficient students speak less, and vice-versa.

4. Conclusion

The key findings from the quantitative analysis seem to be that lexical 'chunks' of data introduced in the focus-on-form stage were reused by students at a rate of around 16% for so-called 'significant' 4- and 5- word strings and that re-use of 67 4+-word strings was detected. It also appears likely that the number of strings of 4+-words reused is an indicator of some form of acquisition. However, this raises several questions, primarily whether 16% is a high or low rate of re-use and whether 67 strings reused constitute a successful performance or not. It is not possible to provide anything other than tentative responses to such questions as no other study (to my knowledge) has attempted to measure acquisition in this way, for this approach, and in such a context. Studies on 'retention' cite differing figures, depending on measurements, type of approach and context, and each have their own definition of 're-use' or 'retention'. For example, Loewen's (2005) study on retention from incidental focus-on-form reported rates of 50% two weeks later, while Bruton (2010) cites a study in which zero retention occurred. In the present study, the only certainty is that some retention did occur, but only another research project conducted under similar (ideally the same) conditions could provide us with a meaningful comparison.

As to the question of what constitutes re-use, it may well be that this study's focus on the reoccurrence of the exact same words has failed to detect re-use in other forms. For example, in this study, we had the reoccurrence of the phrase "(*The Palestinians want) to reduce the size of the Israeli State…*", but what if a student had, instead, said, "*The people of Palestine want a reduction in Israel's size…*"? The words "*the*", "*want*" and "*size*" would appear on both individual word lists, and the lists of family words would contain these three plus "*Palestine*", "*reduce*", and "*Israel*", which would, together, be enough to convey the meaning but, unfortunately, computer software, as it stands, would not be able to produce results with anything other than the words separated on lists. Therefore, unless the human reader of the transcripts has a keen eye and memory, this "reoccurrence" would go undetected.

Regardless of whether such cases constitute a reoccurrence or not, the crucial question here is whether measuring lexical reoccurrence truly equals measuring effectiveness. In answer to this, I would say that this type of analysis, covering individual words to long phrases (in this case, up to eleven words long) can tell us a great deal about the degree of re-use of 'chunks' of language in terms of lexico-grammatical forms and thus whether the focus-on-form contributed to students acquiring these. In other words,

the lexical analysis tells us a lot about how effective the FoF phase was, in this case in a real, classroom-based EAP context, thus achieving the objective of linking classroom activity to student outcomes.

Nevertheless, despite this, there are a number of potentially relevant factors which were not measured in the study and which it would be useful to obtain data on: for example, external factors such as students' study workloads, extrinsic motivation for the tasks (in terms of needs), and the effects of groupwork, background reading and non-verbal information. It would also be useful to have statistics for *Complexity*, *Accuracy* and *Fluency* (Skehan 1998) as indicators of task performance in general. Another interesting avenue for exploration would be the effect of the debates on the content of students' essays. Clearly, to gain a picture of the overall effectiveness, all of these factors and more would need to be measured and analysed, but this is far beyond the scope of one single study which, it is hoped, will form part of the larger jigsaw-puzzle that is to be solved.

Alongside the above recommendations for researchers, concerning the need for a broader range of data from this and similar studies, for both comparison and analysis, I have a number of suggestions for teachers of similar courses, based on the findings of this study.

First of all, I believe that, in such a course as the one mentioned here, two debates are not enough if developing students' language skills in as comprehensive a manner as possible is the goal. Although, clearly, some gains can be made (as demonstrated here), it would be much more beneficial to have a course consisting of a series of debates, ideally on similar topics (but not necessarily the same, and chosen by the students in order to increase motivation), encouraging re-use of the same kind of debating language. As an extension of the work begun here, I would suggest maintaining the structure of the instructional sequence (including native-speaker or C2-level debates as focuses-on-form) but, for example, allowing students to use their 'useful language' documents (as a further focus-on-form) in a second debate, but not in a third. It would be interesting to discover whether such a possibility improved students' language use substantially more than the sequence used in this research. To complement this, based on the high degrees of correlation between the number of chunks retained by individuals and their C-test scores and numbers of words spoken, I would also suggest that teachers 'push' (see Swain 1985) students to use more sophisticated language, to avoid short turns, and encourage the linguistically weaker ones to speak more, through the introduction of clear assessment criteria and feedback, further focuses-on-form, imposition of a more rigid debating structure and increased delegation.

To sum up, in this study, linguistic gains were clearly made through use of a task-based instructional sequence, but I would suggest that, in a future course of this nature, by implementing the above recommendations emerging from this research, further gains in language proficiency can be made from classroom activity using such a task-based approach.

APPENDIX List of new 4+-word strings reoccurring in T2

Strings are listed as *number of words*, *string (frequencies in FoF, T2), speaker*.
(non-contiguous strings plus those within similar native-speaker phrases are highlighted in *italics*). Strings appear ranked in order of i. length (number of words reused), ii. frequency in FoF, iii. frequency in T2 (as a new occurrence), iv. (part-)recurring in FoF (denoted by #), v. alphabetical order. (strings also occurring in T1 but reused by (a) different student(s) are ranked last and denoted by *)

11 What else could (there) *be* (as) *the capital of the new Palestinian state*? (1,1) PD #

10 the Palestinian(s) are very much in favour of/for *a two-state solution* (1,1) PD #
Essentially/basically *we are willing to accept* the/a *Israeli right to exist, but* (1,1) PD

9 parts of Jerusalem where there is an Arab majority (1,1) IF

8 to reduce the size of the Israeli state (1,1) PB

7 we didn't talk about this but (1,1) IF
The U.S. is very much/at first *a pro-Israeli state and* (1,1) PD

6 correct me if I'm wrong (1,1) PO
return back behind/to *the borders of 1967* / the nineteen hundred sixty-seven (1,1) PD
of the State of Israel and (1,1) PB *

5 in the State of Israel (2,2) IF IMI #
to take the first step (2,1) IMI #
what more do you want (2,1) IF #
accept our right to exist (1,2) PD
side would be willing to (1,1) PB #
the/a *two-state solution is* (not) *a solution* (1,1) PD #
there will be no terrorist attacks / *terrorism* (1,1) PO #
things that happened in Europe (1,1) PD #
right to exist but we (1,1) PD
settlements in the west bank (1,1) IMA
the Israelis are willing to (1,1) PB
the Palestinian state-to-be (1,1) PB
the U.S. have a (lesser) *role* (1,1) PNI
we insist/exist *on the fact that* (1,1) PD
in our point of view (erroneous!) (1,2) IF IO *
not be any terrorist attacks (1,1) IN *

4 there has to be (5,2) PB #
the capital of the (3,2) IN PB #
there will be no (3,2) PO IO #
are you talking about (2,3) PO #
has/have to look at PB (2,1) #
going to be a (2,1) IS #
in the Gaza strip (2,1) IS

and the Gaza Strip (1,2) IMI
I would say that (1,2) PD
I think we should (1,2) PB IN
want a peaceful solution (1,2) PB
a lot of things (1,1) IMI
arguing/argue (about the fact) *that you have* a/the *right* (1,1) IF
be/are *willing to try to* (1,1) PO
compromise (which) *could be reached* (1,1) IF
I find the argument (1,1) IF
I mean, there are (1,1) IF
Israelis/Israel had/have a/the *right to defend themselves* (1,1) PB
I think, in this (1,1) PO
Jerusalem is another question (1,1) PNI
need to receive the (1,1) PB
of a Palestinian state (1,1) PB
on the Israeli side (1,1) PB
talking about the (exact) *borders* (1,1) PB
the point of view (1,1) PD
to give up the occupied / these *territories* (1,1) PO
to guarantee that there (1,1) IF
to the Jewish (-Israeli) *state* (1,1) PO
to the Palestinian side (1,1) PB
we can talk about (1,1) IS
from the Palestinian side (2,1) IF *
the fact that the (1,2) PO *
for the Israelis and (1,1) IMI *
if you want to (1,1) IMA *
in the West Bank (1,1) PB *
one of the most (1,1) IF *
representative of the Palestinian(s) (1,1) PD *
their point of view (1,1) IF *
there is a problem (1,1) PO *
what do you think (1,1) IS *
what would you say (1,1) PD *

Bibliography

Alexander, O./Argent, S./Spencer J. (2008): EAP Essentials: A teacher's guide to principles and practice. Reading: Garnet.

Bachman, L. (1990): *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bruton, A. (2002): From tasking purposes to purposing tasks, in: *ELT Journal*, 56, 280-288.

Bruton, A. (2010, April): *Task: a panacea for too much?* Presentation at IATEFL Conference, Harrogate.

Charles, M. (1996): Practice or performance?, in: Hewings, M. and Dudley-Evans, T. (eds): *Evaluation and Course Design in EAP*, Hertfordshire: Prentice Hall Macmillan in association with the British Council.

Council of Europe (2008): *Structured overview of all CEFR scales*, http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Overview_CEFRscales_EN.pdf (25.2.2014)

Ellis, R. (2009): Task-based language teaching: sorting out the misunderstandings, in: *International Journal of Applied Linguistics*, 19, 221-246.

Feez, S. (1998): *Text-based syllabus design*. Sydney, NSW: AMES.

Gillet, A./Wray, L. (2006): EAP and success, in: Gillet, A./Wray, L. (eds): *Assessing the Effectiveness of EAP Programmes*, London: BALEAP.

Krashen, S. (1982): Principles and Practice in Second Language Acquisition. Oxford: Pergamon.

Loewen, S. (2005): Incidental focus on form and second language learning, in: *Studies in Second Language Acquisition*, 27(3), 361–86.

Long, M. (1983): Native speaker/non-native speaker conversation and the negotiation of comprehensible input, in: *Applied Linguistics*, 4 (2), 126−41.

Long, M./ Porter, P. (1985): Group work, interlanguage talk, and second language acquisition, in: *TESOL Quarterly*, 19, 207-228.

Lynch, B. (1996): *Language Program Evaluation*. Cambridge: Cambridge University Press.

Richards, J.C./Rodgers, T.S. (2001): *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Schmidt, R. (1990): The Role of Consciousness in Second Language Learning. London: Edward Arnold.

Skehan, P. (1998): *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, P. (2009): Modelling second language performance: integrating complexity, accuracy, fluency and lexis, in: *Applied Linguistics*, 30/4, 510-532.

Storch, N./Tapper, J. (2009): The impact of an EAP course on postgraduate writing, in: *Journal of English for Academic Purposes*, 8, 207-223.

Swain, M. (1985): Communicative competence: Some roles of comprehensible input and comprehensible output in its development, in: Gass, S.M. /Madden C.G. (eds): *Input in second language* acquisition, Rowley, MA: Newbury House.

Swan, M. (2005): Legislation by hypothesis: The case of task-based instruction, in: Oxford: *Applied Linguistics*, 26, 376-401.

Williams, J. (2001): The effectiveness of spontaneous attention to form, in: *System*, 29, 325- 340.

Mat Plews M.A., Sprachenzentrum, Humboldt-Universität zu Berlin,
Unter den Linden 6, 10099 Berlin, matthew.plews@spz.hu-berlin.de