References

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education: Principles, Policy and Practice, 14*(2), 201–232.

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice, 18*(3), 259–78.

Baker, W.H. (2001). HATS: A design procedure for routine business documents. *Business Communication Quarterly, 64*(2), 66–78.

Beddow, P.A., Elliott, S.N., & Kettler, R.J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International, 2013*, Article ID 952704, 1–12.

Beddow, P.A., Kurz, A., & Frey, J.R. (2011). Accessibility theory: Guiding the science and practice of test item design with the test-taker in mind. In S.N. Elliott, R.J. Kettler, P.A. Beddow, & A. Kurz (Eds), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice and policy* (pp.163–182). New York: Springer.

Chelesnik, A.L. (2009). *The impact of test typography on student achievement, anxiety, and typographic preferences*. Unpublished Masters Thesis. California State University, San Marcos. Available from: https://csusm-dspace.calstate.edu/handle/10211.3/114688 (retrieved 4th November 2019).

Crisp, V. (2011). Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties. *Irish Educational Studies, 30*(3), 323–343.

Crisp, V., Johnson, M., & Novaković, N. (2012). The effects of features of examination questions on the performance of students with dyslexia. *British Educational Research Journal, 38*(5), 813–839.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48*(2), 139–154.

Evett, L., & Brown, D. (2005). Text formats and web design for visually impaired and dyslexic readers – Clear Text for All. *Interacting with Computers, 17*(4), 453–472.

Ketterlin-Geller, L.R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3–16.

Kettler, R.J., Dickenson, T.S., Bennett, H.L., Morgan, G.B., Gilmore, J.A. et al. (2012). Enhancing the accessibility of high school science tests: A multistate experiment. *Exceptional Children, 79*(1), 91–106.

Lonsdale, M. dos S., Dyson, M. C., & Reynolds, L. (2006). Reading in examination-type situations: The effect of text layout on performance. *Journal of Research in Reading, 29*(4), 433–453.

Moncada, S.M., & Moncada, T.P. (2010). Assessing student learning with conventional multiple-choice exams: Design and implementation considerations for business faculty. *International Journal of Education Research, 5*(2), 15–29.

OCR. (2018a). GCSE (9–1) *Gateway Science: Exploring our question papers*. Cambridge: OCR. Retrieved from: https://www.ocr.org.uk/Images/462559-exploring-our-question-papers-gateway-science.pdf (retrieved 4th November 2019).

OCR. (2018b). GCSE (9–1) *Twenty First Century Science: Exploring our question papers*. Cambridge: OCR. Retrieved from: https://www.ocr.org.uk/Images/462607-exploring-our-question-papers-twenty-first-century-science.pdf (retrieved 4th November 2019).

OECD. (2009). *Learning mathematics for life: A perspective from PISA (Programme for International Student Assessment)*. Paris: OECD Publishing.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards* (pp.166–206). London: Qualifications and Curriculum Authority.

QCA. (2004). *The statutory regulation of external qualifications*. London: Qualifications and Curriculum Authority.

QCA. (2005). *Fair access by design: Guidance for awarding bodies and regulatory authorities on designing inclusive GCSE and GCE qualifications*. London: Qualifications and Curriculum Authority.

Spalding, V. (2009). *Is an examination paper greater than the sum of its parts? A literature review of question paper structure and presentation*. Manchester: AQA Centre for Education Research and Policy.

Using corpus linguistics tools to identify instances of low linguistic accessibility in tests

David Beauchamp and **Filio Constantinou** Research Division

Introduction

Assessment is a useful process as it provides teachers and other stakeholders (e.g., parents, government, employers) with information about students' competence in a particular subject area. However, for the information generated by assessment to be useful, it needs to support valid inferences. One factor that can undermine the validity of inferences from assessment outcomes is the language of the assessment material. For instance, if a Mathematics test question contains complex vocabulary and/or grammar, it might prevent students from demonstrating their true mathematical knowledge and skills. This may result in teachers and other stakeholders drawing inaccurate inferences from the test scores. Students who are not native speakers of the target language are more likely to be disadvantaged by assessment material that displays low levels of linguistic accessibility. In an attempt to support teachers and test developers in designing linguistically accessible assessment material, this study explored practical ways of investigating the complexity of test questions both at the level of vocabulary (lexical complexity) and grammar (syntactic complexity).

The starting point of this research was the shortcomings of traditional measures of linguistic accessibility, or readability, and their limited applicability to test questions. For example, traditional readability measures often assume that longer words are more difficult to comprehend (see Lenzner, 2014). However, in the context of assessment, such words are normally subject-specific technical terms (e.g., *microorganism, photosynthesis*) with which students are expected to be familiar, as they are part of the construct that is being assessed. Also, traditional readability measures tend to be based upon continuous prose and fully formed sentences and, as a result, are not well-suited for measuring the readability of texts that do not fit this format, especially multiple-choice questions for example. Furthermore, readability measures that are based on sentence length and text length do not consider the different cognitive challenges that various syntactic structures pose on readers (Lenzner, 2014).

In response to these shortcomings, alternative ways of investigating the linguistic accessibility of assessment materials were explored. These involved undertaking lexical and syntactic analyses of test questions in an automated manner using software packages typically employed in the field of corpus linguistics (for a definition of corpus linguistics, see the Method section below). To our knowledge, this study represents one of the first attempts to identify instances of low linguistic accessibility in assessment material using corpus linguistics methods. In this study, accessibility is understood as "the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge of the target construct [and] is conceptualized as the sum of interactions between features of the test and individual test taker characteristics" (Beddow, Elliott, & Kettler, 2013, p.1).

Lexical complexity

The issue of lexical complexity, or lexical sophistication, in testing is often discussed in the context of the assessment needs of second language speakers. Second language speakers constitute a particularly vulnerable group as they are assessed via a language that is different from their mother tongue. In the context of high-stakes testing, characteristic is the study by Shaw and Imam (2013) that sought to identify, among other linguistic resources, the vocabulary needed by non-native English speakers to complete IGCSEs in History, Biology and Geography successfully. The lexical resources were then classed according to the Common European Framework of Reference for Languages (CEFR), an international scale that describes language competence (Council of Europe, 2018).

An important distinction to make when considering the challenge that vocabulary poses to a test taker is that between *content-obligatory language* and *content-compatible language* (Cloud, Genesee, & Hamayan, 2000). The former includes technical, subject-specific language needed to understand and respond to test items (e.g., *photosynthesis* and *Reformation* for Biology and History respectively), while the latter is a foundation of more common, non-subject-specific language (e.g., *plants* and *social development* for Biology and History respectively). This distinction is important because when identifying instances of lexical complexity which may compromise accessibility, one must discount what is likely to be content-obligatory vocabulary, the learning and use of which makes up part of the construct to be assessed.

Research concerned with the lexical complexity of texts has involved the compilation of vocabulary level lists that have been used in lexical analysis software, such as the RANGE program and AntWordprofiler, and in tests designed to assess learners' lexical knowledge such as the Vocabulary Size Test (see Anthony, 2013; Bauer & Nation, 1993; Beglar & Nation, 2007; Nation, 2018; Webb & Nation, 2008). The most extensive vocabulary level lists are based upon language use in the British National Corpus (BNC) and Corpus of Contemporary American English (COCA) (see Nation, 2018). Each level of these lists consists of vocabulary derived from 1000 word families (a word family is vocabulary based around a root word such as *give*, and its derivatives such as *gives*, *giving*, *given*). In particular, Level 1 consists of vocabulary based upon the first 1000 word families an English language learner is likely to encounter, Level 2 is based upon the next thousand word families, and so on. The vocabulary grows progressively more obscure through 29 levels. Table 1 below provides examples of words across the different levels, as found in the BNC/COCA vocabulary lists and in A Level Biology examination papers.

Table 1: Examples of words across levels, as found in the BNC/COCA vocabulary lists and in A Level Biology examination papers

Vocabulary list level	*Examples from the BNC/COCA vocabulary lists*	*Examples from A Level Biology examination papers*
1	offer (offers, offered); stay (stays, stayed, staying); carry (carries, carried, carrier)	what; show; main; that; student
2	access (accesses, accessed, accessible); fry (fries, fried, fryer)	section; indicator; repeated
3	abandon (abandons, abandoned, abandoning); collapse (collapses, collapsed, collapsing); promote (promotes, promoted, promoters)	vessel; theory; evolved
4	abnormal (abnormality, abnormalities, abnormally); prestige (prestiges, prestigious); subsidiary (subsidiaries, subsidiarity)	graph; acid; interval
5	accessory (accessorise, accessorised, accessories); burgle (burgled, burglar, burglaries); lurk (lurks, lurked, lurking)	saturate; niche; botany
6	abduct (abducted, abducting, abduction); clutter (clutters, cluttered, cluttering); incubate (incubates, incubated, incubation)	chromosome; receptor; aquatic
7	abate (abated, abatement, abating); ludicrous (ludicrously, ludicrousness); throng (throngs, thronged, thronging)	tentacle; amphibian; viral
8	abstinence (abstinences); orator (oratories, orators, oratory); paraphrase(paraphrases, paraphrased, paraphrasing)	catalyse; yoga; biodiversity
9	abyss (abysses, abyssal); denominator (denominators)	photosynthesis; collagen; microorganism
10+	adage (adages); libertine (libertines); portcullis	habituate; hydrolysis; glycaemic

Syntactic complexity

Syntactic complexity is concerned with linguistic structures above the level of the individual word (e.g., clauses, sentences). Syntactically complex texts can increase cognitive load and thus undermine accessibility by placing the barrier of good reading skills and good working memory before the construct to be tested.

Research into linguistic accessibility has identified syntactic features that can affect comprehension. Štajner, Evans, Orasan and Mitkov (2012) reported that subordinating phrases, coordinating phrases, infinitives and prepositional phrases as grammatical structures were generally associated with a lower degree of readability on the Flesch scale (see Table 2 for examples of some of these structures). In a similar vein, Ariel (2001) has developed a spectrum of linguistic accessibility markers which ranges from *low accessibility* markers (e.g., long descriptions, long noun phrases) to *high accessibility* markers (e.g., pronouns, noun omission), with less linguistic material generally being more favourable for cognitive processing. The level of conceptual content within sentences has also been considered as a factor that may affect readability. For instance, Feng, Jansche, Huenerfauth and Ehladad (2010) found that the number of general nouns and named entities in a text, also known as entity-density, performed well as a readability measure, with greater *entity-density* indicating lower readability.

Subordinating phrases in the form of nested clauses (clauses embedded within other clauses) are considered to increase linguistic complexity, as they require greater mental effort on the part of the reader to be successfully processed (Gibson, 1998; Miller & Isard, 1964). However, it is not only the presence of certain syntactic features that can affect the complexity of a sentence. The position of such features within the sentence can also have implications for complexity and, by extension, accessibility. In a study on survey question difficulty, Lenzner (2014) points to the difficulty in processing left-branching structures. These are structures which contain considerable linguistic material in the form of clauses, phrases or other modifiers before the main verb is reached. The need to process this linguistic material prior to encountering the main verb in the sentence tends to increase the demand on working memory. The sentences below exemplify the difference between left-branching (a) and right-branching (b) structures:

> *(a) How likely is it that if a law was considered by parliament that you believed to be unjust or harmful, you, acting alone or together with others, would try to do something against it?*
>
> *(b) How likely is it that you, acting alone or together with others, would try to do something against a law that was considered by parliament and that you believed to be unjust or harmful?*
>
> Lenzner (2014, p.685)

Concerning examination papers, much work has focused on the presence of linguistic complexity in Mathematics papers, probably owing to the risk that excessive language processing poses to an assessment in which the target construct is essentially a non-linguistic one. A range of studies have been carried out examining the effects of aspects of syntactic complexity on the performance of EAL (English as an Additional Language) students in Mathematics tests, using candidate interviews, DIF (Differential Item Functioning) statistics and regression analyses (e.g., Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Wolf et al., 2008). These studies focused on various markers of syntactic complexity in Mathematics papers including sentence length, item length, noun phrase length, and the presence of prepositional phrases, participles and multiple and relative clauses. Their results showed that the effects of syntactic complexity on candidate performance were limited or inconclusive. Table 2 below illustrates how these syntactic features are manifested in A Level Biology examination papers.

Table 2: Features contributing to syntactic complexity as manifested in A Level Biology examination papers

Syntactic feature	*Example*
Subordinating clause	Complete Table 1 by putting a tick in a box **if the structure is present in the type of cell**. The reserve managers chose a high temperature **because this causes the young lizards to hatch more quickly**. **Although a moss plant has no vascular tissue,** water still moves through the plant from the root-like structures to the leaves.
Passive structure	The volunteers **were asked** to record three symptoms.
Prepositional adjunct	The circles in Figure 1 represent the hierarchy **of taxonomic groups for the classification** shown **in Table 1**.
"to" infinitive	He used a pH meter **to record** pH.
Past participle phrase	The table below shows the vitamin C content of sauerkraut and cabbage, **treated in different ways**.
Present participle phrase	**Using a genetic diagram**, find the probability that the next child born to parents 3 and 4 would be affected by moyamoya.
Relative clause	The photograph below shows packaging pellets made from thermoplastic starch, **which is produced from corn starch**. This investigation was carried out in a university laboratory, using species of bacteria **that cause disease in humans.**

With a view to making mathematics items more linguistically accessible to candidates, Abedi and Lord (2001) simplified verb phrases, conditional clauses, relative clauses, question phrases and abstract representations. They found that EAL and non-EAL students alike made small but statistically significant improvements on simplified items, as did students from low socio-economic backgrounds. Additionally, they found that items that had been simplified were more likely to be selected by candidates when a choice was given.

Method

To investigate linguistic accessibility in assessment material, three corpora of examination papers were compiled. The examination papers were obtained from three A Level subjects that represented different disciplines: Biology, Business Studies and History. The papers were developed by three major examination boards in England and were taken by students in the UK between 2015 and 2017. Each corpus was approximately 15,000 words long and comprised several hundred examination questions, covering a wide range of examples of examination questions typically encountered by candidates. The three corpora were explored using software packages commonly employed in corpus linguistic studies.

Corpus linguistics can be defined as a method of analysing "the actual patterns of use in natural texts" (Biber, Douglas, Conrad, & Reppen, 2004, p.4). It involves compiling large bodies of text, or corpora, and analysing them via specialist software to identify the presence, distribution and frequencies of various linguistic features. Analysing language use by means of corpus linguistics software, rather than manually, has certain advantages. These include (a) the capacity to analyse large amounts of text within a very short amount of time, and (b) the ability to identify trends that may be missed through an 'intuitive' reading by an individual. To our knowledge, to date, corpus linguistics software has not been used to investigate language use in assessment materials.

In this study, two corpus linguistics software packages were mainly used: AntWordProfiler (Anthony, 2013) and Multidimensional Analysis Tagger (Nini, 2015). The former was used for the lexical analysis, while the latter was used for the syntactic analysis.

AntWordProfiler: lexical analysis

AntWordProfiler is a software program which allows corpora of texts to be compared to imported word lists (Anthony, 2013). The software ranks the words in the texts according to their level of complexity (i.e., the inferred likelihood of a person knowing a word based upon the frequency of its use within a corpus of real language use). In this study, the BNC/COCA vocabulary level lists (see Nation, 2018) were used to provide a scale against which the vocabulary in the examination papers could be ranked. More specialised and technical vocabulary (e.g., scientific and historical terms) forms the content of the higher lists, while more commonplace, non-technical vocabulary forms the content of the lower lists. To frame these lists in a more widely known scale, Nation has provided an approximate classification of these vocabulary level lists based on the CEFR levels via personal communication (P. Nation, 21 September, 2018). This approximate classification is shown in Table 3 below.

As can be seen in Table 3, vocabulary which is present in lists 5 to 9 may not be known by candidates who are not "proficient" in English. As such, it could be viewed as representing a barrier to accessibility. On the other hand, vocabulary found in lists 10 and higher tends to be specialist or technical vocabulary that forms part of content-obligatory language and, as such, it is likely that it will have been encountered by candidates. However, it should be noted that this is not always the case. For instance, as can be seen in Table 1 above, there are examples of technical terms which are found in lists lower than level 10 (e.g., 'photosynthesis' which appears in list 9).

Multidimensional Analysis Tagger: syntactic analysis

Multidimensional Analysis Tagger (MAT) is a software package that analyses plain text files and uses a parts-of-speech (POS) tagger to identify and label syntactic features (Nini, 2015). The results of the analyses are then displayed in a table format. From these results, it is possible to isolate the presence and frequency of relevant syntactic features and structures across different texts. The syntactic features considered in this study are shown in Table 4 (see also Nini, 2015). They were chosen because: (a) they represent multiword structures which increase the linguistic material (and thus cognitive load) of the text; (b) they represent a variety of different semantic relations between entities; and (c) some of them have been shown in previous studies to affect text readability (see e.g., Štajner et al., 2012). The chosen features are not necessarily considered to be equal in the challenges they pose to readability.

It should be noted that MAT does not carry out the syntactic analysis at the level of the sentence or the clause, but only at the level of the provided text file. In this study, some of the analyses were carried out at the level of the subject corpus, while some others were carried out at the level of the item. Although the syntactic features considered in this study may have indicated the presence of syntactic complexity, the way in which the complexity was distributed among different sentences had to be identified through manual human analysis.

Table 3: Classification of BNC/COCA vocabulary level lists based on CEFR

CEFR level			*BNC/COCA vocabulary level lists*	
Proficient	C2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.	7000–9000 words	Lists 7–9+
	C1	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.	5000–6000 words	Lists 5–6
Independent	B2	Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.	4000 words (2000–3000 high frequency words plus 1000–2000 relevant technical vocabulary)	List 4
	B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.	2000–3000 high frequency words	Lists 2–3
Basic	A2	Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.	The most frequent 1000 word families	List 1
	A1	Has a basic vocabulary repertoire.	120 words and phrases from the survival vocabulary (=vocabulary needed for coping with simple survival needs)	List 1

Table 4: Syntactic features considered in this study (see Nini, 2015)

	Syntactic feature	*Example*
Additional clauses, beyond the simple structure of subject + verb + object	Causative adverbial clauses indicated by *because*.	*The business failed* ***because there was a lack of demand for the product***.
	Concessive adverbial clauses indicated by the words *although* and *though*.	***Although the snakes are venomous***, *they rarely approach humans*.
	Conditional adverbial clauses indicated by the words *if* and *unless*.	*The campaign would be more successful* ***if it used targeted advertising***.
	Other adverbial subordinating clauses signalled by words such as *since*, *while* and *whereas*.	***Whereas the economy of the Northern states was increasingly industrial***, *the economy of the Southern states remained predominantly agricultural*.
Passive structures		The journal ***is published*** biannually by the press.
Prepositional adjuncts		***In 1871***, *Germany was unified* ***by Bismarck***.
"to" infinitives		*They agreed* ***to stop*** *selling the product after the lawsuit*.
Participles: past and present		***Built in a single week***, *the house would stand for fifty years.* ***Stuffing his mouth with cookies***, *Joe ran out of the door*.
Relative clauses	*Pied-piping* relative clauses: any preposition followed by *who*, *whom*, *whose*, or *which*.	*The manner* ***in which he was told***.
	That relative clauses in an object position.	*The dog* ***that I saw***.
	That relative clauses in a subject position.	*The dog* ***that bit me***.
	Sentence relatives: indicated by a punctuation mark followed by *which*.	*Bob likes fried mangoes,* ***which is disgusting***.
	What clauses	*I believed* ***what he told me***.
	Who relative clauses in an object position.	*The man* ***who Sally likes***.
	Who relative clauses in a subject position.	*The man* ***who likes popcorn***.

Findings

Key observations from the lexical and syntactic analyses are presented below.

Lexical analysis

There was variation in the level of lexical complexity that was observed across the three corpora. While the vast majority of vocabulary was indicated to be at an accessible level (89.7%–93.6% of vocabulary lay within levels 1 to 4), each subject corpus included examples of vocabulary of increasing complexity and obscurity which could potentially disadvantage some candidates, especially EAL ones. The Biology corpus displayed the highest proportion of language at levels 5 to 9 (4.3% as opposed to 1.3% in Business Studies and 1.2% in History), while the History corpus displayed the highest proportion of vocabulary at levels 10 and above. As mentioned earlier, according to Nation, vocabulary at levels 5 to 9 tends to correspond to vocabulary expected of second language speakers who are at CEFR levels C1 and C2 (i.e., "proficient level"). On the other hand, vocabulary at level 10 and above often indicates subject-specific vocabulary, or content-obligatory language (including proper nouns and dates). Characteristic examples of words which may disadvantage EAL students (i.e., words which are not subject-specific and are at level 5 or above) can be found in the following History items:

> *Assess which religious issue most* ***hindered*** *the development of […] in the period from […].*
>
> *Study all the Sources. Use your own knowledge to assess how far the Sources support the interpretation that the difficulty in finding a solution to the problems of […] was the* ***reluctance*** *of the […] to co-operate with […].*

More examples of non-subject-specific vocabulary at levels 5 to 9 that occurred in the examination papers analysed can be found in Table 5.

Table 5: Non-subject-specific vocabulary at levels 5 to 9 used in examination papers in Biology, Business Studies and History

Level	*Biology*
Level 5	*miniature, voyage, expel*
Level 6	*stranded, streamline*
Level 7	*rupture, tar, deduce*
Level 8	*dissociate, frill*
Level 9	*sheath*
Level	*Business Studies*
Level 5	*incur, mattress, morale, pier, ruthless, hawk, trailer, flop, goose, grooming, ignite, orphan, underestimate, abolish, brochure*
Level 6	*souvenir, outweigh, stout, hygiene, drawback, wasp, glossy, mentor*
Level 7	*bingo, scaffold, titan, ware*
Level 8	*gourmet, posh, fang, aptitude*
Level 9	*fizz, kiln*
Level	*History*
Level 5	*hinder, voyage, reluctance*
Level 6	*influx*
Level 7	*blunder, gravely, misplace*
Level 8	*hermit*
Level 9	–

Where there was uncertainty as to whether certain words were subject-specific or not, AntConc (a free concordancing and text analytics package) (see Anthony, 2018) was used to identify occurrences of these words in the respective syllabi. However, it should be acknowledged that the distinction between 'subject-specific' and 'non-subject-specific' vocabulary is not clear-cut and that some non-subject-specific words that are relatively rare in everyday discourse may also be encountered in specific classroom teaching (e.g., 'tar' in the context of health risks of smoking).

Syntactic analysis

As the analyses carried out via MAT showed, the three subject corpora displayed considerable differences in terms of their use of grammatical features that tend to contribute to syntactic complexity. For example, "to" infinitives were comparatively over-represented in Business Studies, suggesting a focus on verbs and actions. Similarly, passives predominated in Biology, suggesting a tendency towards more formal language and the reporting of processes. Although these observations indicate little in terms of the accessibility of individual items, they suggest differences in item construction across subjects.

Pairs of items which were similar in some respects (e.g., were obtained from the same subject; were of similar length) but had a comparatively high or low frequency of the target syntactic features were closely examined. The aim of this more fine-grained analysis was to identify how these features manifested themselves in the context of the items and whether they posed a threat to accessibility. Two such items are presented and discussed below. The items, which were of similar length (Item 1: 46 words; Item 2: 49 words), were obtained from Biology examination papers. The frequency of the target syntactic features for Item 1 and Item 2 can be found in Tables 6 and 7 respectively.

> ***Item 1:***
>
> *Hormonal control of […] is achieved by hormones acting on the […]. Using your knowledge of the way in which […] is coordinated, suggest why it can be deduced that hormones act on the […] rather than on individual […] cells.*

Table 6: Item 1: Target syntactic features per 100 tokens (as generated by MAT)

			Item 1			
Tokens	*Additional clauses*	*Passives*	*Prepositional adjuncts*	*"to" infinitives*	*Participles*	*Relative clauses*
46	0	4.35	17.39	0	4.34	4.34

> ***Item 2:***
>
> *The table below shows the mean […] rate and the standard deviation (SD) for the […] treatment group and the control group. Plot a suitable graph to show all the data for the […] treatment group. Do not include the standard deviations. Join the points with ruled, straight lines.*

Table 7: Item 2: Target syntactic features per 100 tokens (as generated by MAT)

			Item 2			
Tokens	*Additional clauses*	*Passives*	*Prepositional adjuncts*	*"to" infinitives*	*Participles*	*Relative clauses*
49	0	0	6.12	2.04	0	0

Item 1 can be described as a more complex text. The second sentence contains a present participle ("*Using*…") that modifies the main command verb "*suggest*", instructing students on what to do to answer the question. In addition, there are two nested clauses ("…*the way in which*…" and "…*why it can be deduced*…") and two passive structures ("…*is achieved by*…" and "…*is coordinated*…") which amount to 31 words, as well as multiple entities that need to be processed by the candidate.

In contrast, Item 2 comprises four short sentences, none exceeding 21 words, with simple subject-verb-object (sentence 1) and imperative-object structures (sentences 2, 3 and 4). The four sentences have mostly short noun phrases, contain minimal extra information in the form of prepositional adjuncts and no nested relative clauses. Also, there is no preceding modification of the main command verbs "*plot*" and "*join*".

Overall, of these two similarly sized items, Item 1 appears less accessible due to its longer sentences, its greater number of nested structures and its lengthy, left-branching participle leading up to the main verb in the second sentence ("*Using your knowledge of*…, *suggest*…" which requires the candidate to process additional linguistic material before reaching the main verb of the sentence).

Discussion

This study compiled three corpora of examination papers and used corpus linguistics techniques to explore linguistic accessibility in examination questions. The lexical and syntactic analyses to which the corpora were subjected, via AntWordprofiler and MAT respectively identified trends that invite closer attention.

AntWordProfiler, when used in conjunction with the vocabulary level lists, can help to identify low-frequency vocabulary that may inhibit reading comprehension, especially for candidates who do not have English as a first language. Vocabulary which does not represent content-obligatory language but is categorised above level 4 (i.e., it is at "proficient level" according to CEFR) might be considered complex and likely to introduce construct-irrelevant variance into test scores. When such vocabulary is identified by software and judged by question writers to be indeed complex, alternatives should be sought. A comparison of synonyms against the vocabulary level lists could help question writers to identify more accessible lexical substitutes. For instance, in the examples above, 'obstructed' could be used in place of 'hindered' (Level 5 vocabulary), while 'unwillingness' or 'hesitation' could be used in place of 'reluctance' (Level 5 vocabulary). Even though some words appear less sophisticated and therefore more accessible than others, it would be useful for future research to attempt to evaluate the effect of lexical substitutions on candidates' performance. Such evaluations may help to provide not only a more empirical basis for the need to exhibit lexical sensitivity in item writing but also indicate the forms that such lexical sensitivity should take in practice.

With respect to syntactic complexity, software such as MAT can be used to profile syntactically individual items and identify the frequency of features that could influence syntactic complexity. The qualitative comparison of pairs of more and less syntactically complex items of similar length may help to identify linguistic structures and item writing styles likely to prove barriers to accessibility. As shown in this study, examples of such linguistic structures and/or styles include left-branching constructions (signalled by features such as participles), the presence of multiple entities to be processed (signalled by features such as prepositional adjuncts), longer sentences (signalled by features such as additional clauses and prepositional adjuncts), and multiple and nested clauses (signalled by features such as relative pronouns and subordinating conjunctions). Where relatively inordinately high levels of such features are found in items, the items could be flagged for further consideration and potentially for revision to improve accessibility. As the examples examined in this study indicated, items that displayed a higher concentration of these features appeared to be less accessible than similarly sized items that displayed a lower concentration of the target features. However, to enable the automated identification of excessively complex items in the future, further research is required. Such research can draw on developments in the field of linguistics and test in an experimental manner the accessibility of different linguistic configurations of items to help identify empirically-derived principles of linguistic accessibility.

In conclusion, corpus linguistics tools have not been typically used in item writing. However, as this study has demonstrated, they can prove particularly useful by providing directions for the improvement of items. Apart from helping to identify items that may display low levels of linguistic accessibility, they can also be used as training instruments in professional development courses intended for prospective as well as experienced item writers. Arguably, corpus linguistics tools can help to raise awareness among item writers of the ways in which different linguistic features and different item writing styles can hinder or enable the measurement of students' true abilities.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Anthony, L. (2013). AntWordProfiler (Version 1.4.0w) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from: http://www.laurenceanthony.net/software

Anthony, L. (2018). AntConc (Version 3.5.2) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from: http://www.laurenceanthony.net/software

Ariel, M. (2001). Accessibility theory: an overview. In T. Sanders, J. Schilperood & W. Spooren (Eds), *Text representation: linguistic and psycholinguistic aspects* (pp.29–87). Amsterdam: John Benjamins.

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: item reviews and lessons learned from four state assessments. *Education Research International, 2013*.

Beglar, D., & Nation, I.S.P. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Biber, D., Douglas, B., Conrad, S., & Reppen, R. (2004). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Cloud, N., Genesee, F., & Hamayan, E. (2000). *Dual language instruction: a handbook for enriched education*. Boston, MA: Heinle & Heinle.

Council of Europe (2018). *Use of the CEFR*. Retrieved from: https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp.276–284). Association for Computational Linguistics.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76.

Lenzner, T. (2014). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods and Research, 43*(4), 677–698.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333–368.

Miller, G. A., & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control, 7*(3), 292–303.

Nation, I.S.P. (2018). *The BNC/COCA word family lists*. Retrieved from: https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf

Nini, A. (2015). Multidimensional Analysis Tagger (Version 1.3). Retrieved from: http://www.academia.edu/4285869/Multidimensional_Analysis_Tagger_v_1.3

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.

Shaw, S., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly, 10*(4), 452–475.

Štajner, S., Evans, R., Orasan, C., & Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (pp.14–22).

Webb, S., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, 16, 1–10.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., Bachman, P. L., Chang, S. M, Farnsworth, T., Jung, H., Nollner, J., & Shin, H. W. (2008). Providing validity evidence to improve the assessment of English language learners. CRESST Report 738. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.