45

Corpora in language testing: developments, challenges and opportunities

Sara T. Cushing

1 What is language testing?

This chapter provides an overview of how corpus linguistics tools and techniques can be used in language testing and assessment (LTA). The terms testing and assessment are often used interchangeably, although testing is more often associated with large-scale, high-stakes examinations, while assessment is associated with evaluation in the classroom. For the purposes of this chapter, I will use both terms to mean the same thing, though for the most part I will be discussing large-scale testing.

LTA has been defined as 'the systematic gathering of language-related behavior in order to make inferences about language ability and capacity for language use on other occasions' (Chapelle and Plakans 2012: 241). Given this definition, it is not surprising that there are numerous applications of corpus linguistics to the enterprise of language testing, since the systematic collection and analysis of language data are at the very heart of corpus linguistics. What makes LTA distinct from other applications of corpus linguistics is the second half of the definition: 'to make inferences about language ability and capacity for language use on other occasions'. That is, while corpus linguists strive to collect representative existing examples of language in a particular context, language tests are intended to elicit novel instances of language performance or other observable language-related behaviour, from which inferences about underlying ability or predictions of future behaviour are made (Xi 2017).

More importantly, the inferences made from language test scores typically serve as the basis, alone or in combination with other data, for making decisions about people or programmes. These decisions may have high or low stakes for an individual test taker. Low-stakes decisions based on test scores include such things as whether a student decides to put more effort into learning or a teacher decides that their class is ready to move on to a new lesson. High-stakes decisions are much more consequential, such as whether a student will be admitted to a university, an airline pilot will be allowed to fly on international routes or a family will be allowed to immigrate to a new country. Thus, high-stakes tests frequently serve as gatekeeping devices in education, immigration and employment contexts (McNamara *et al.* 2019). For this reason, language testing scholars

DOI: 10.4324/9780367076399-45

are concerned with the fairness of these decisions and with ensuring that the information provided by a test is sufficient and appropriate to the decision being made. In other words, validity in language testing is a central concern, as a test claims to provide evidence of the communicative ability of a person that is sufficient for making decisions, and the nature of this evidence must be scrutinised to ensure the fairness of decisions based on the test.

Before considering how corpora can be used in language testing, it may be useful to define the scope of LTA and outline the steps involved in creating a language test for a given purpose. Useful introductions to LTA can be found in numerous sources, such as Bachman and Palmer (1996), Green (2013) and Fulcher (2013), to name a few.

The process of test development is iterative and recursive, in that the phases of test development are not always distinct and may be completed concurrently or revisited at a later time, but for the sake of simplicity they are presented here in a linear fashion. Once the purpose for the test has been established, test development first involves defining the construct, or ability, to be measured. The construct typically includes both the nature of language to be elicited and the context in which the language will be used: e.g. reading and writing English for academic purposes or spoken Japanese for tour guides. The next step is to conduct an analysis of what is often called the Target Language Use (TLU) domain (Bachman and Palmer 1996), which involves systematically collecting and analysing authentic examples of the language used in the setting for which the test takers will need to use the language in terms of lexis, grammatical structures and typical language functions. Test developers next need to make decisions about what task types will be included and write specifications, or blueprints, for individual test items or task types and for the test as a whole. Once these preliminary decisions are made, items are written, reviewed and revised as necessary and answer keys, rubrics or scoring guides are written. Items are pilot tested so that their statistical characteristics are known (i.e. item difficulty and discrimination, or the degree to which an item can distinguish between lower- and higher-ability test takers) and to make sure that items and directions are clear. Pilot testing may result in additional modifications to test tasks, rubrics or individual items. Decisions also need to be made about how scores will be used (i.e. cut-off scores for particular decisions) and reported to test takers and other test users. Once the test development team is satisfied with the results of the pilot testing, the finalised test is assembled, administered and scored and the results disseminated, with ongoing monitoring for quality control.

Traditionally, tests have been evaluated based on at least four essential qualities: *reliability*, or consistency across test items, forms, and raters; *validity*, or evidence that the test is measuring what it claims to measure; *practicality*, the capability of administering and scoring a test given limited resources of time, money, and person hours; and *washback*, or beneficial consequences of the test (see, for example, Green 2013, for more information). As noted earlier, validity is often seen as the central concern of testing, so a more in-depth discussion of validity may be useful here.

A traditional view of validity involves what is sometimes called the "three Cs" of validity: content, criterion and construct validity. Evidence for content validity – the match of the test content to the skills or abilities being tested—typically comes from expert review of the test items to make sure that they test the intended skills and knowledge and that they do so as comprehensively as possible within the constraints of the test. Criterion-related validity has to do with the relationship between performance on the test and performance on another measure of the same ability, either administered

at the same time (concurrent validity) or at some future date (predictive validity). For example, if I wanted to demonstrate that my 15-minute test of English was useful for university admissions, I could administer it along with TOEFL or IELTS and calculate the correlation between these two tests (concurrent validity), or I could calculate the correlation between my test and students' grade point averages at the end of their first semester of study (predictive validity). Construct validity has to do with the relationship between test performance and a theoretical construct. Models such as Bachman and Palmer's model of communicative language ability (Bachman and Palmer 1996) have been developed to provide the theoretical underpinnings for language tests. Since Messick's (1989) seminal paper on validity, LTA scholars have come to see construct validity as all-encompassing; that is, the evidence that a test is measuring what it claims to be measuring can include both content and criterion-related evidence, as well as evidence that relates the test more directly to a theoretical construct, often through sophisticated statistical means.

More recently, scholars in LTA have been using argument-based frameworks for test validation based on the writing of Michael Kane (e.g. 1992, 2013). Kane expands upon Messick (1989), agreeing with Messick that it is not the test itself that is valid, but the 'adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment' (1989: 13). Kane's contribution to the field is to lay out in greater detail the assumption and warrants that link test performance to these inferences and actions in what he calls an interpretive argument. The evidence supporting these warrants, or claims, comprises a validity argument.

Other scholars in language testing have expanded upon Kane's notion of an interpretive argument; Bachman and Palmer (2010), for example, lay out what they call an Assessment Use Argument for an assessment that involves several claims. Similarly, Chapelle *et al.* (2008) present an interpretive argument for the TOEFL iBT comprising six key inferences. While a complete discussion of these inferences is beyond the scope of this chapter, I will return to them later in discussing the ones that are most relevant to corpus linguistics.

To summarise, research in LTA comprises efforts not only to develop useful assessments that are fit for a specific purpose but to provide evidence supporting a chain of inferences that lead from test performance (observed behaviour) to decisions that are made on the basis of that performance and the consequences of those decisions. In the next section I provide some historical background on the application of corpus linguistics to these endeavours.

2 Milestones in the use of corpora in language testing

Prior to the 1990s, while corpus linguistics was beginning to make inroads in the related fields of second language studies and language teaching, most large-scale testing programmes did not rely on corpus data to inform test development or validation (Barker 2006). It was not until 1996 that a prominent LTA scholar, Charles Alderson, laid out a case for the usefulness of corpus data for test development (Alderson 1996: 254). One of the first authors to publish an article describing the use of corpus data in test development was Coniam (1997), who proposed a method for using corpus-based word frequency data for the automatic generation of cloze tests. Meanwhile, corpus linguists within applied linguistics were beginning to point out the usefulness of corpus data to LTA specialists around the same time; see, for example, Biber *et al.* (1998).

Since 2000, more attention has been paid to the use of corpus data in LTA, with review articles written by Barker (2006, 2010, 2013), Taylor and Barker (2008), Park (2014), Weigle and Goodwin (2016) and Cushing (forthcoming). A symposium on corpora in language testing was organised at the Language Testing Research Colloquium, the major international LTA conference, in 2003 (Taylor and Barker 2008), and in 2017, a special issue of *Language Testing* was devoted to applications of corpus linguistics to LTA (Cushing 2017). A recent review of ten years of language testing research (Plakans 2018: 6) notes that the topic of technology in testing journals had become a "deluge", with corpus linguistics and automated scoring (which relies heavily on corpus data) as the two most important areas within technology discussed.

The increased interest in corpus linguistics within LTA comes from both the increased availability of appropriate corpora and the development of new tools for analysing corpus data that are accessible to researchers with limited programming knowledge. Another driver for the need to use corpus data in assessment is the proliferation of automated scoring engines for writing, and increasingly for speaking as well. In the rest of the section I outline these considerations.

LTA researchers and test developers draw on both learner corpora, as a basis for examining the features of language used by learners at different levels of proficiency, and specialised reference corpora, so that test content can be reflective of authentic language use in the TLU domain. An excellent survey of early corpora of interest to language testing is found in Taylor and Barker (2008). In the early 1990s, researchers under the direction of Sylviane Granger began to compile the *International Corpus of Learner English* (ICLE), which consists of the writing of advanced learners of English, along with background information of the writers such as age, first language and so on (see Granger *et al.* 2002) (see Chapters 22 and 23, this volume). Another important learner corpus, the *Cambridge Learner Corpus* (CLC), developed by the EFL Division of the University of Cambridge Local Examinations Syndicate (UCLES) in association with Cambridge University Press, consists of exam scripts (i.e. written responses) from Cambridge examinations at different levels of proficiency, along with demographic information about the candidates. A proportion of the CLC is error-coded, allowing for searches of specific errors alongside more traditional lexical and collocational searches (Taylor and Barker 2008)

Around the same time, test developers began compiling or commissioning specialised corpora specifically to inform test development. An important example is the *TOEFL 2000 Spoken and Written Academic Language Corpus* (T2K-SWAL), a large corpus of spoken and written academic language commissioned by the Educational Testing Service (ETS), to inform the transformation of the paper-based TOEFL to what is now the TOEFL internet-based test (TOEFL iBT; Biber *et al.* 2004). Similarly, the *Pearson International Corpus of Academic English* (PICAE; Ackermann *et al.* 2011) was compiled to inform the development of the Pearson Test of English Academic, comprising written and spoken curricular and extracurricular materials.

A milestone event that has had a major impact on LTA was the publication of the Common European Framework of Reference (CEFR, Council of Europe 2001), originally intended to provide a common vocabulary for referring to proficiency levels across European languages. The six CEFR levels range from A1 (beginner) to C2 (mastery), with descriptions of what learners can do at each level of proficiency. Since its publication, many test developers, particularly in the UK and Europe, have aligned their tests to the CEFR, so that a passing score on a particular test is evidence that a

candidate has reached a given level, such as B2 or C1. The CEFR can-do statements are functional rather than linguistic in nature, so scholars have long been interested in exploring the characteristics of the language used at different levels of proficiency. For example, one of the can-do statements at Level A2 is that learners 'can communicate in simple and routine tasks'; researchers are interested in knowing what vocabulary and language structures (e.g. verb tenses, clause structure and so on) learners have control over to successfully communicate in routine situations. Test developers and curriculum designers also became interested in specifying the characteristics of texts that would serve as appropriate input for each level. The English Profile programme (Hawkins and Buttery 2010; Hawkins and Filipović 2012) is an example of a large project intended to identify "criterial features" for each CEFR level using texts from the CLC; i.e. features that will distinguish learners at one level from those at adjacent levels (Salamoura and Saville 2010) (see also Chapter 22, this volume).

As corpora began to be used more for assessment, corpus analysis tools along with natural language processing tools were improving as well, lending themselves to applications to LTA. Before the late 1990s, corpus tools tended to be limited in their functionality, including only being able to process ASCII characters (Anthony 2013). Tools such as *AntConc* and *Wordsmith* made corpus data more widely available, as they could be used on larger corpora, were not limited to ASCII characters, included commonly used statistics and, importantly, had user-friendly interfaces (see Chapter 9, this volume). In recent years, computational linguists have begun developing free online tools that can provide sophisticated analyses of such phenomena as cohesion, lexical sophistication and syntactic complexity, which are of particular relevance to language testing scholars seeking to understand the linguistic characteristics of texts at various levels of proficiency. Coh-Metrix (Graesser *et al.* 2004) and Lu's Syntactic Complexity Analyzer (Lu 2010) are examples.

An important development in language assessment has been the increased use of computer-delivered tests, which allow for the possibility of automated scoring and feedback on language production, particularly for writing. These systems rely heavily on corpus data, i.e. large collections of essays that have been scored by trained raters. Automated scoring systems typically predict human scores by measuring a set of features in the texts to be scored that are considered relevant to the construct, even if they may not be identical to the features human raters attend to in scoring (Williamson *et al.* 2012). Even though automated scoring was originally developed in the 1960s, it has only been applied to large-scale language tests in the past two decades or so (see Dikli 2006, for a review). One of the first automated scoring systems to be used on high-stakes tests was e-rater, developed by ETS and first used on the GMAT in 1999 (Williamson *et al.* 2012). E-rater is currently used in conjunction with human scorers on the TOEFL iBT Independent and Integrated writing tasks. The Pearson Test of Academic English was the first large-scale test to feature completely automated scoring, using the Intelligent Essay Assessor (IEA) scoring engine by Knowledge Analysis Technologies (Landauer *et al.* 2003). A relatively new addition to the landscape is the Duolingo Test of English (LaFlair and Settles 2019), which is administered online and is automatically scored. In any of these systems, natural language processing techniques allow for the extraction of numerous linguistic features from a corpus of texts and then determining the combination of features that best predict human scores.

Similarly, automated feedback systems such as Cambridge English's *Write and Improve* (https://writeandimprove.com/) rely on large learner corpora as a basis for

provide instant feedback on writing, including highlighting possible sentence-level errors; see Stevenson and Phakiti (2019) for a recent review of the potential and current limitations of automated feedback systems. Automated scoring of speaking is less advanced than scoring of writing but similarly relies on large corpora of learner data. As anyone who uses a smart phone knows, current voice-to-text software relying on automated speech recognition (ASR) is not completely reliable, and is even less so for learners of a language, due to such things as deviations from the norm in grammar, vocabulary and pronunciation, along with an increase in disfluencies and hesitations (Litman *et al.* 2018). In assessment, such systems are currently only reliable for highly constrained tasks such as reading aloud or providing highly predictable responses to questions. Improvements in ASR for language learners will only improve with better corpora of learner data for training these systems.

3 Using corpora for test development

Corpus linguistics can inform both test development and test validation. For test development, corpus data are useful both for targeting test items at particular proficiency levels and for ensuring that test language is both appropriate for the test audience and authentic to a particular language use situation (see discussion of content validity earlier). In testing, a distinction can be made between selected response items (sometimes called objective items), such as multiple-choice or matching items, and constructed response items (or subjective items), such as prompts for speaking or writing. Selected response items are typically, though not exclusively, used for assessing the receptive skills of reading and listening and for assessing discrete enabling skills such as grammar and vocabulary. Both reference corpora and learner corpora are useful for developing these items. As Barker (2010, 2013) notes, corpus data can help item writers base their work on authentic language and target specific aspects of language that are relevant to a population of test takers. For example, learner corpus data can provide insights into collocational patterns typical of learners at different proficiency levels, which can help test developers identify collocations to include in a test targeted at a particular level (see Voss 2012, as an example) and also identify collocational errors that may be useful as distractors.

Corpus data can confirm or disconfirm intuitive judgments about language use patterns, which were the basis for much test design in the mid-to-late twentieth century. Word frequency is often thought to be a useful proxy for difficulty, for example. However, Alderson (2007: 402) found that correlations between expert judgments and objective frequency of words in the *British National Corpus* were 'only moderate'.

Corpus data are particularly relevant to specific purpose testing; that is, tests that are used to certify language proficiency for specific occupations, which may have their own specialised lexicon or other structures. An example of a corpus designed for a specific purpose is discussed in Moder and Halleck (2012), who compiled a corpus of authentic communications between pilots and air traffic controllers, which served as the basis for the Versant Aviation English Test (Van Moere *et al.* 2009)

Constructed responses tend to be lengthier instances of language production in speaking or writing based on a prompt. In independent production tasks, the prompt is generally simple (no more than a sentence or two). In integrated tasks, test-takers must first read and/or listen to one or more input texts and then respond to a prompt that requires incorporating content from the input. While it is more common for test

developers to rely on corpus data for developing listening and reading items, Xi (2017) suggests that corpus analysis can be useful in developing language tasks for speaking and writing as well, based on the analysis of a reference corpus, which might suggest that certain linguistic features are salient in a given language use context that is relevant to an assessment. Aspects of the task, such as specification of the purpose or audience for the language production in the prompt, might be manipulated to elicit those salient features more naturally. Another application of corpus data is ensuring comparable 'opportunity of use' (Caines and Buttery 2018: 6) that is, test takers need to have equal opportunities to display their language competence. In a study of responses to a variety of writing prompts in the CLC, they found that different task-topic types tend to elicit different lexico-syntactic constructions, suggesting that test developers should control for these factors in task development.

Extended speaking or writing tasks are typically evaluated by human raters using a rubric or rating scale that defines several levels of performance in terms of their communicative effectiveness and/or the characteristics of the language used; computers are increasingly being called upon to automate scoring, to be discussed later. Corpus data can be useful for developing or improving these scales. For example, Hawkey and Barker (2004) developed a common set of writing descriptors that could be applied across tests at different proficiency levels using corpus analysis techniques. Römer (2017) uses corpus linguistics tools to argue against the traditional separation of lexis and grammar in rating scales in favor of multi-word expressions that cross the boundaries between grammar and vocabulary, suggesting that rating scales for spoken language might benefit from considering lexico-grammatical ability as a single construct.

4 Using corpora for test validation

As noted earlier, many LTA scholars have promoted an argument-based approach to test validity in which the inferences leading from the test performance to the consequences of decisions based on test scores are outlined and then evidence gathered to either support or refute these inferences. At least three of these inferences can be supported using corpus data: domain description, explanation and extrapolation. The inference of domain description in a test states that test performance reveals skills, knowledge and abilities that are representative of the target domain (Chapelle *et al.* 2011) – that is, where the language will ultimately be used, such as academic settings, employment, etc. Corpus evidence for this inference comes from an analysis of the situational characteristics of the target domain along with the lexico-grammatical characteristics of language used in the domain. Examples of corpus linguistics methods being used to support domain description include several studies using the T2KSWAL corpus to analyse college-level academic spoken and written language (e.g. Biber *et al.* 2004; Biber and Gray 2013) and the use of corpus-based techniques using a systemic functional linguistics perspective to compare the knowledge structures and language functions of international teaching assistants (ITAs) with those found in a corpus of TOEFL iBT speaking test responses (Cotos and Chung 2018).

The explanation inference connects test scores to a theoretical construct or latent ability being measured by the test. Corpus-based studies of test-taker language production in constructed responses and their relationships to scores support the explanation inference. In terms of validating scoring rubrics, Knoch and Chapelle (2018: 489) present the following warrant supporting the explanation inference: 'the descriptors

in the rating scale…are identifiable in the candidates' discourse in the response'. Evidence for this warrant can come from a corpus-based analysis of candidate discourse in reference to the scale descriptors and the features of discourse that differentiate between adjacent scoring levels. Examples of corpus-based studies that address this inference include Banerjee *et al.* (2007), Cumming *et al.* (2005) and Friginal and Weigle (2014).

Finally, the extrapolation argument states that performance on a test is related to performance in the target domain. One study explicitly using corpus data to investigate the extrapolation inference is LaFlair and Staples (2017), who conducted a corpus-based register analysis to compare the linguistic features elicited by an oral proficiency interview with the language features of several registers in the TLU domains for the test, including both academic and nursing contexts. Evidence that the language elicited by test tasks approximates language expected in the target domain provides support for the extrapolation inference.

In summary, corpus analyses are useful for test validation in three principal ways: First, comparing test features with appropriate reference corpora provides evidence that the test content adequately represents the domain of interest; second, corpus analysis of test-taker responses with respect to both the descriptors in the rating scale and investigations of the features of test-taker language that are related to test scores provides evidence that test performance is related to the construct of interest; and finally, comparisons of test-taker language with language produced in authentic real-life situations provide evidence that test performance is related to performance in relevant non-test situations.

5 Caveats and future developments

It is clear from this discussion that corpus linguistics has much to offer LTA. Before moving on to possibilities for the future, however, I will briefly discuss cautions and caveats that have been raised with regard to the use of corpus data to inform language test development and validation. The first caveat has to do with the choice of corpora to represent the target domain, input to learners or learner production. Egbert (2017) cautions against conflating learner corpora with corpora of test responses, since test responses tend to fall near the unnatural end of a continuum of naturally occurring language. The degree to which test responses produced under timed conditions for the purpose of displaying language ability are comparable to L2 language produced outside of a testing situation for authentic communication is an empirical question, relating to the extrapolation inference referenced earlier. Another issue that scholars need to contend with is the appropriateness of using L1 corpora to represent the target language domain. Egbert suggests that it is particularly problematic to use L1 corpora as a proxy for input or the exposure of L2 learners to the language. For example, he critiques the use of COCA as a proxy for L2 learner experience in Kyle and Crossley's (2017) study of verb–argument constructions, since the registers in COCA are unlikely to be identical to the registers that learners will encounter, nor does the balance of texts among different registers in COCA replicate the experience of learners.

Both Egbert (2017) and Xi (2017) caution against too much reliance on computer programs that extract, count and analyse linguistic features from corpora in test validation or in the development of automated scoring engines, since the linguistic features that can be extracted automatically are not always relevant to the construct being

assessed. For such analysis to be useful in language testing, the extracted features must be related to a construct with a clear operational definition. Similarly, both authors express caution over the direct application of findings from corpus analysis to scoring rubrics. As Egbert (2017: 563) notes, a corpus analysis can reveal patterns of language use that may not be salient to human raters, and it is necessary to establish whether raters can be trained to notice and reliably assess these patterns.

Furthermore, Xi (2017:571) notes that the corpus linguist's goal of providing rich description of authentic language used in naturalistic environments does not completely overlap with the goal of test developers, which is to 'provide an adequate representation of the target language use domain and to elicit a performance that is indicative of a potential performance in corresponding real-world contexts', rather than 'to mimic faithfully real-world tasks or elicit exactly the same language as in the target domain'. Language testers have developed frameworks and tools to select a subset of relevant linguistic features and test tasks that are feasible and practical in a given context, and it is unrealistic to expect a test to replicate the conditions of authentic language use.

Despite these caveats, there are numerous ways in which new developments in corpus linguistics can inform LTA and developments in LTA can benefit from corpus data. In this section, drawing primarily on recommendations by Park (2014), Xi (2017) and Weigle and Goodwin (2016), I discuss some of these developments. First, there is a need for large longitudinal learner corpora, preferably with error annotations and parallel corpora of corrections. Such corpora would be useful to trace language development over time and across proficiency levels, assist in accurately identifying criterial features that are reliably associated with different proficiency levels and mitigate the serious challenge of establishing comparability between learner and expert corpora. These corpora could also be useful in formative and dynamic assessment, where learners could have access immediately to exemplars of authentic language appropriate for their current stage of learning (see also Chapters 22 and 23, this volume).

In addition, corpora of regional varieties of English, such as English used in Singapore or Malaysia, would also be valuable in establishing the legitimacy of these varieties in the local context. Speakers of these varieties are often assessed against criteria representing a more prestigious variety of English (e.g. British or American English), which may be less relevant in the local contexts (Park 2014).

In terms of corpus analysis tools, as natural language processing (NLP) tools and knowledge become more sophisticated, it would be helpful for language testers to have more user-friendly corpus-based tools at their disposal. A suite of freely available text analysis tools by Kris Kyle and Scott Crossley is available at https://www.linguisticanalysistools.org/. These are welcome additions to the language tester's toolkit, with the caveats mentioned earlier. Tools for the automated analysis of features beyond syntax and vocabulary, such as textual cohesion and organisation, are critically important, as are tools that can extract a wider range of phenomena found in spoken language (Xi 2017).

One area where corpus linguistics is providing new insights is in assessing interactional competence, defined both as a psychological (individual) construct and as a social construct, in that interaction is co-constructed between participants (McNamara and Roever 2006). Galaczi and Taylor (2018) identify several contributions of corpus linguistics to understanding the features of interaction, such as the role of turn-opening tokens (e.g. Tao 2003) or the use of stance markers by L2 speakers (Gablasova *et al.* 2015) (see also Chapters 22 and 23, this volume). This is an area that is ripe for future development.

Another area where work is beginning to appear is in adapting spoken dialog systems (SDSs) for assessment use to simulate the one-on-one interaction of an oral interview in proficiency tests (Litman *et al.* 2018). While such systems are in common use in applications such as call centers and intelligent personal assistants such as Apple's *Siri* or Amazon's *Alexa*, the corpus data needed for such an effort would have to go beyond what is currently available for SDS, such as dialogs including data from non-native speakers that have been assessed for proficiency (see Litman *et al.* 2016, for an example).

For more progress in these areas, I concur with Xi's (2017) call for broader cooperation between corpus linguists, computational linguists and assessment scholars. As Xi points out, such collaboration can lead to advances in the following areas: (1) improving automated scoring systems by incorporating more construct-relevant features of language, (2) enhancing automated feedback systems by making feedback more meaningful to teaching and learning and (3) using automated linguistic analysis tools to develop learning progressions.

Further reading

Barker, F. (2013) 'Using Corpora to Design Assessment', *The Companion to Language Assessment* 2: 1013–28. (This article provides information on the use of corpora in large-scale test design.)

Cushing, S. T. (ed.) (2017) 'Corpus Linguistics in Language Testing Research' [Special issue], *Language Testing* 34(4). (This special issue presents five articles that use corpus linguistics tools and techniques to explore LTA issues, along with an introduction by the editor and commentaries by a corpus linguist and an LTA specialist.)

Park, K. (2014) 'Corpora and Language Assessment: The State of the Art', *Language Assessment Quarterly* 11(1): 27–44. (This article provides an overview of computational approaches to language assessment and advances in the use of corpora for LTA.)

References

Ackermann, K., De Jong, J. H. A. L., Kilgarriff, A. and Tugwell, D. (2011) 'The Pearson International Corpus of Academic English (PICAE)', *Proceedings of Corpus Linguistics*, Available at: https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-47.pdf.

Alderson, J. C. (1996) 'Do Corpora Have a Role in Language Assessment?' in J. Thomas and M. Short (eds), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, London: Longman, pp. 248–59.

Alderson, J. C. (2007) 'Judging the Frequency of English Words', *Applied Linguistics* 28(3): 383–409.

Anthony, L. (2013) 'A Critical Look at Software Tools in Corpus Linguistics', *Linguistic Research* 30(2): 141–61.

Bachman, L. F. and Palmer, A. S., (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (2010) *Language Assessment in Practice: Developing Language Assessments and Justifying their Use in the Real World*, Oxford: Oxford University Press.

Banerjee, J., Franceschina, F. and Smith, A. M. (2007) 'Documenting Features of Written Language Production Typical at Different IELTS Band Score Levels', *International English Language Testing System (IELTS) Research Reports 2007* 7(1). Canberra: IELTS Australia and British Council.

Barker, F. (2006) 'Corpora and Language Assessment: Trends and Prospects', *Research Notes* 26: 2–4.

Barker, F. (2010) 'How can Corpora be Used in Language Testing?' in A. O'Keeffe and M. J. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, 1st edn, London: Routledge, pp. 661–74.

Barker, F. (2013) 'Using Corpora to Design Assessment', in A. Kunnan (ed.) *The Companion to Language Assessment*, Vol. 2, Oxford: John Wiley and Sons, pp. 1013–28.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E. and Urzua, A. (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus (TOEFL Monograph Series MS-25)*. Princeton, NJ: Educational Testing Service.

Biber, D. and Gray, B. (2013) *Discourse Characteristics of Writing and Speaking Task Types on the TOEFL ibt® Test: A Lexico-Grammatical Analysis (ETS Research Report Series, 2013(1)*. Princeton, NJ: Educational Testing Service.

Caines, A. and Buttery, P. (2018) 'The Effect of Task and Topic on Opportunity of Use in Learner Corpora', in V. Brezina and L. Flowerdew (eds) *Learner Corpus Research: New Perspectives and Applications*, London: Bloomsbury, pp. 5–27.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (eds) (2008) *Building a validity argument for the Test of English as a Foreign Language™*, London: Routledge.

Chapelle, C. A. and Plakans, L. (2012) 'Assessment and Testing: Overview', in C. A. Chapelle (ed.) *The Encyclopedia of Applied Linguistics*, Hoboken, NJ: John Wiley and Sons, Inc, pp. 241–44.

Coniam, D. (1997) 'A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests', *Calico Journal* 14: 15–33.

Cotos, E. and Chung, Y. R. (2018) *Domain Description: Validating the Interpretation of the TOEFL iBT® Speaking Scores for International Teaching Assistant Screening and Certification Purposes (ETS Research Report Series RR-18-45)*, Princeton, NJ: Educational Testing Service.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U. and James, M. (2005) *Analysis of Discourse Features and Verification of Scoring Levels for Independent And Integrated Prototype Writing Tasks for New TOEFL (TOEFL Monograph No. MS-30)*, Princeton, NJ: Educational Testing Service.

Cushing, S. T. (2017) 'Corpus Linguistics in Language Testing Research', *Language Testing* 34(4): 441–9.

Cushing, S. T. (forthcoming) 'Corpus Linguistics and Language Testing', in G. Fulcher and L. Harding (eds) *The Routledge Handbook of Language Testing*, 2nd edn, London: Routledge.

Dikli, S. (2006) 'An Overview of Automated Scoring of Essays', *The Journal of Technology, Learning and Assessment* 5(1): 1–36

Egbert, J. (2017) 'Corpus Linguistics and Language Testing: Navigating Uncharted Waters', *Language Testing* 34(4): 555–64.

Fulcher, G. (2013) *Practical Language Testing*, London: Routledge.

Friginal, E. and Weigle, S. C. (2014) 'Exploring Multiple Profiles of L2 Writing Using Multidimensional Analysis', *Journal of Second Language Writing* 26: 80–95.

Gablasova, D., Brezina, V., McEnery, T. and Boyd, E. (2015) 'Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style', *Applied Linguistics* 38(5): 613–37.

Galaczi, E. and Taylor, L. (2018) 'Interactional Competence: Conceptualisations, Operationalisations, and Outstanding Questions', *Language Assessment Quarterly* 15(3): 219–36.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004) 'Coh-Metrix: Analysis of Text on Cohesion and Language', *Behavior Research Methods, Instruments, and Computers* 36(2): 193–202.

Granger, S., Dagneaux, E. and Meunier, F. (eds) (2002) *The International Corpus of Learner English. Handbook and CD-ROM*, Louvain-la-Neuve: Presses Universitaires de Louvain.

Green, A. (2013) *Exploring Language Assessment and Testing: Language in Action*, London: Routledge.

Hawkey, R. and Barker, F. (2004) 'Developing a Common Scale for the Assessment of Writing', *Assessing Writing* 9(2):122–59.

Hawkins, J. and Buttery, P. (2010) 'Criterial Features in Learner Corpora: Theory and Illustrations', *English Profile Journal* 1: E5. doi: 10.1017/S2041536210000103

Hawkins, J. A. and Filipović, L. (2012) *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, Vol. 1, Cambridge: Cambridge University Press.

Kane, M. T. (1992) 'An Argument-Based Approach to Validity', *Psychological Bulletin* 112: 527–35.

Kane, M. T. (2013) 'Validating the Interpretations and Uses of Test Scores', *Journal of Educational Measurement* 50: 1–73.

Knoch, U. and Chapelle, C. A. (2018) 'Validation of Rating Processes within an Argument-Based Framework', *Language Testing* 35(4): 477–99.

Kyle, K. and Crossley, S. (2017) 'Assessing Syntactic Sophistication in L2 Writing: A Usage-Based Approach', *Language Testing* 34(4): 513–35.

LaFlair, G. T. and Staples, S. (2017) 'Using Corpus Linguistics to Examine the Extrapolation Inference in the Validity Argument for a High-Stakes Speaking Assessment', *Language Testing* 34(4): 451–75.

LaFlair, G. T. and Settles, B. (2019) *Duolingo English Test: Technical Manual*, Pittsburgh, PA: Duolingo, Retrieved 6/9/2020 from https://s3.amazonaws.com/duolingo-papers/other/Duolingo%20English%20Test%20-%20Technical%20Manual%202019.pdf.

Landauer, T. K., Laham, D. and Foltz, P. W. (2003) 'Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor', in M. D. Shermis and J. C. Burstein (eds) *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 87–112.

Litman, D., Strik, H. and Lim, G. S. (2018) 'Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities', *Language Assessment Quarterly* 15(3): 294–309.

Litman, D., Young, S., Gales, M., Knill, K., Ottewell, K., van Dalen, R. and Vandyke, D. (2016) 'Towards Using Conversations with Spoken Dialogue Systems in the Automated Assessment of Non-Native Speakers of English', in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 270–75.

Lu, X. (2010) 'Automatic Analysis of Syntactic Complexity in Second Language Writing', *International Journal of Corpus Linguistics* 15(4): 474–96.

McNamara, T., Knoch, U. and Fan, J. (2019) *Fairness, Justice and Language Assessment*, Oxford: Oxford University Press.

McNamara, T. and Roever, C. (2006) *Language Testing: The Social Dimension*, Malden, MA/ Oxford, UK: Blackwell.

Messick, S. (1989) ' Validity', in R. L. Linn (ed.) *Educational Measurement*, 3rd edn, Washington, DC: American Council on Education and National Council on Measurement in Education, pp. 13–103.

Moder, C. L. and Halleck, G. B. (2012) 'Designing Language Tests for Specific Social Uses', in G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing*, London: Routledge, pp. 137–49.

Park, K. (2014) 'Corpora and Language Assessment: The State of the Art', *Language Assessment Quarterly* 11(1): 27–44.

Plakans, L. (2018) 'Then and Now: Themes in Language Assessment Research', *Language Education and Assessment* 1(1): 3–8.

Römer, U. (2017) 'Language Assessment and the Inseparability of Lexis and Grammar: Focus on the Construct of Speaking', *Language Testing* 34(4): 477–92.

Salamoura, A. and Saville, N. (2010) 'Exemplifying the CEFR: Criterial Features of Written Learner English from the English Profile Programme', *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research. EuroSLA Monographs Series* (1): 101–32.

Stevenson, M. and Phakiti, A. (2019) 'Automated Feedback and Second Language Writing', in K. Hyland and F. Hyland (eds) *Feedback in Second Language Writing: Contexts and Issues*, Cambridge: Cambridge University Press, pp. 125–42.

Tao, H. (2003) 'Turn Initiators in Spoken English: A Corpus-Based Approach to Interaction and Grammar', in P. Leistyna and C. F. Meyer (eds) *Corpus Analysis: Language Structure and Language Use*, Amsterdam: Rodopi, pp. 187–207.

Taylor, L. and Barker, F. (2008) 'Using Corpora in Language Assessment', in E. Shohamy and N. Hornberger (eds) *Language Testing and Assessment*, *Encyclopedia of Language and Education*, Vol. 7, New York: Springer, pp. 241–54.

Van Moere, A., Suzuki, M., Downey, R. and Cheng, J. (2009) 'Implementing ICAO Language Proficiency Requirements in the Versant Aviation English Test', *Australian Review of Applied Linguistics*, 32(3): 1–17.

Voss, E. (2012) 'A Validity Argument for Score Meaning of a Computer-Based ESL Academic Collocational Ability Test Based on a Corpus-Driven Approach to Test Design', unpublished PhD dissertation, Lowa State University.

Weigle, S. C. and Goodwin, S. (2016) 'Applications of Corpus Linguistics in Language Assessment', in J. V. Banerjee and D. Tsagari (eds) *Contemporary Second Language Assessment: Contemporary Applied Linguistics*, Vol. 4, New York, NY: Bloomsbury, pp. 209–24.

Williamson, D. M., Xi, X. and Breyer, F. J. (2012) 'A Framework for Evaluation and Use of Automated Scoring', *Educational Measurement: Issues and Practice* 31(1): 2–13.

Xi, X. (2017) 'What Does Corpus Linguistics Have to Offer to Language Assessment?' *Language Testing* 34(4): 565–77.