Designing a South African Multilingual Learner Corpus of Academic Texts (SAMuLCAT)

Adelia Carstens
https://orcid.org/0000-0003-1518-6170
University of Pretoria, South Africa
adelia.carstens@up.ac.za

Roald Eiselen
https://orcid.org/0000-0002-8612-5175
South African Centre for Digital Language Resources (SADiLaR), North-West University, South Africa
roald.eiselen@nwu.ac.za

Abstract

This article provides an overview of the process and initial outcomes of designing a multilingual corpus of academic texts produced by university students with different mother tongues in South Africa, with a view to making it available as an open resource for pedagogical applications and research. We first give an overview of the history of corpus development for pedagogical purposes world-wide, with particular emphasis on learner corpora, and highlight the absence of a South African corpus of academic learner texts. Thereafter, the objectives of the corpus project are outlined. The remainder of the article describes and justifies the design-features of the corpus as well as the process of setting up the data management system to facilitate the collection of the learner texts and their integration with the metadata. We conclude with a summary of the current status of the project, including the limitations, and a preview of the way forward.

Keywords: academic texts; learner corpus; multilingual corpus; South African corpus; learner corpus design

1. Introduction

According to the canonical definition of a learner corpus, it is a collection of machine-readable authentic texts (including transcripts of spoken data), produced by learners of a second or foreign language, which is representative of a particular language or language variety (McEnery, Xiao and Tono 2006, 5; Granger 2012, 3235; Gilquin 2015, 10). Recently, the focus of learner corpora has been broadened in two dimensions. First, the notion of a "learner" has been extended from individuals learning a second or a foreign language to individuals acquiring a new register, such as the academic register. In this broader sense, the British Academic Written English (*BAWE*) corpus, which is a collection of texts produced by undergraduate and Master's students in the United Kingdom in a wide range of disciplines, of which the majority are L1 speakers (Alsop and Nesi 2009, 71), may also be regarded as a learner corpus. Second, a single corpus may be focused on more than one language, in that it has been generated by learners with various first languages, and/or the corpus may include more than one target language. An example is the *MiLC* corpus, which involves the written work of students learning English, Spanish, French and German as foreign languages as well as Catalan as a first, second or foreign language (Römer 2012, 167). Learner corpora have become useful repositories in supporting research on language learning, for instance through systematic error analysis, determining differences between native and non-native language, describing the features of interlanguages, applying learner-corpora research to language teaching methodology and course design (Tono 2003, 804; Granger 2015), and underpinning computer-assisted language learning (CALL) software (Granger 2015).

In South Africa, the interest of a number of researchers working in the domain of Academic Literacy (more detail is given below) has recently been sparked in compiling a comprehensive corpus of academic learner texts. It was deemed that such a corpus would assist in increasing the use of data-driven decisions and designs in curriculum development, materials development and testing. This is further highlighted by the South African Department of Higher Education and Training's repeated call for increasing multilingualism at South African universities and also for developing the indigenous South African languages as languages of higher learning (Department of Higher Education and Training 2017, 2–3). This task does not only require big data on students' use of the current media of instruction but also extensive data on the academic use of the indigenous African languages.

The present article describes the process of designing a *South African Multilingual Learner Corpus of Academic Texts* (*SAMuLCAT*) (in which both "multilingual" and "learner" should be interpreted in their inclusive senses) as an open resource for pedagogical applications and research. We first trace the history of corpus development for pedagogical purposes globally, with particular emphasis on learner corpora, and highlight the limited number and scope of uniquely South African corpora of academic learner texts. This is followed by a description of the specific objectives of the project, the design-features considered during the planning of this project, and setting up the data management system to facilitate the collection of the learner corpus texts and integration of the metadata. The article is concluded by a summary of the current status of the project, its limitations, and a preview of the way forward.

2. Context and Rationale

Since the 1990s there has been a growing interest in applying the findings of corpus linguistics in language pedagogy (Huang 2017, 3). Stimuli in the application of corpora in language teaching include the increased use of computers, the concomitant access to large collections of spoken and written texts in electronic format, globalisation, the development of concordance software for data-analysis using keywords or user attributes, the development of annotation software, and new theories of language and language learning (Richards and Rodgers 2014).

The rapprochement between corpus linguistics and language pedagogy gave rise to the introduction of a unique genre in corpus development, viz. learner corpora, of which the largest are the *International Corpus of Learner English* (*ICLE*) (3.5 million words of learner essays); the *Cambridge Learner Corpus (CLC)* (40 million words); and the *Longman's Learner Corpus (LLC)* (12 million words) (Huang 2017, 6). Both the *CLC* and the *LLC* contain data from compositions written by L2 learners with different first languages. However, they are not available for research; their use is restricted to the publishing houses who own the dictionaries based on these corpora (Pearson-Longman and Cambridge University Press, respectively) (Lozano and Mendikoetxea 2013, 71). The first version of the *ICLE* (Granger, Dagneaux and Meunier 2002) sparked growing interest in learner corpus research and was the first learner corpus to be exploited for pedagogical purposes on a large scale. More recently, established learner corpora that could be described as "large," or are aiming to become large, include the *NUS* (National University of Singapore) *Corpus of Learner English* (*NUCLE*) (Dahlmeier, Ng and Wu 2013), which is freely available for research purposes, and the *CEDEL2* corpus (Lozano and Mendikoetxea 2013).

A comprehensive list of learner corpora around the world (henceforth LCW list) is available on the website of the Centre for English Corpus Linguistics (n.d.). The list contains the names of the corpora, the target language(s), first language(s), medium (spoken or written), task/text type, proficiency level and size in words.

As the LCW list bears witness, no comprehensive learner corpora are available for national or regional varieties of South African English, or for any of the other official South African languages, let alone for academic texts produced by learners (in the narrow as well as the broad sense). The few existing corpora of learner English that originated in South Africa have largely been collected for very specific research purposes. According to B. van Rooy (E-mail correspondence with Tobie van Dyk, May 11, 2017) South African learner corpora include a Tswana Learner English Corpus, comprising 500 essays (compiled by him between 2001 and 2003), available as part of *ICLE*; a corpus built from data collected at the former Vista University (Port Elizabeth campus) and the University of the Witwatersrand; and a corpus of 200 essays written for the National Senior Certificate Examination, collected at the North-West University. In addition to these, one of Van Rooy's PhD students compiled a corpus of 400 argumentative essays by Nigerian learners of English in 2005, which is built on the *ICLE* model.

For Afrikaans, a number of small to medium-sized learner corpora exist, which, similar to the English learner corpora compiled in South Africa thus far, also originated in personal research projects. However, they differ from the above-mentioned English corpora in that the authors were primarily L1 speakers. The Afrikaans learner corpora include Meintjes' corpus of 731 argumentative essays (437 580 tokens) by first-year students who were enrolled for Academic Literacy at the North-West University in 2010 (Z. Meintjes, unpublished notes on the Afrikaans corpus compiled for her PhD study, 2017); Van Rooy's corpus of 60 argumentative essays modelled on the *ICLE* template; and Prinsloo, Taljard and Bosman's corpus of assignments (1.3 million tokens) written by Education students at the University of Pretoria (E. Taljard, pers. comm., May 28, 2018).

3. Objectives

SAMuLCAT was born out of initiatives by board members of ICELDA (the Inter-University Consortium for Language Development and Assessment) as well as out of the interest of other academics at the North-West University and the University of Pretoria. These scholars reiterated the need for a large corpus of South African learner texts to support language planning and development in South African higher education. A funding opportunity became available through the South African Centre for Digital Language Resources (SADiLaR[1]), a government-funded entity of which the overarching aim is to establish "a new language resource infrastructure particularly focusing on languages spoken in Southern Africa, but with an eventual aim to become a hub for digital language resources within Sub-Saharan Africa" (Roux 2016). One of the two main goals of SADiLaR is to provide support for projects implementing digital language-based data (Roux 2016), which also applies to the current project.

The main objective of *SAMuLCAT* is to collect data from various universities in South Africa in order to build a multi-L1, multi-L2, multimodal, multi-genre and multi-level corpus of academic learner texts on the basis of internationally accepted corpus-building principles. Furthermore, the corpus will include metadata about the nature of the task and the characteristics of the individual authors to maximise the utility of the data for researchers. The corpus will be made available as an open resource for basic and applied research, as well as for application to the benefit of South African learners. It is foreseen that the corpus, along with the requisite metadata, will be made available for download from the SADiLaR repository[2] under a Creative Commons Attribution-Noncommercial-ShareAlike 4.0 International license (CC- BY-NC-SA). In addition, it is envisaged that

[1] https://www.sadilar.org

[2] https://repo.sadilar.org

the corpus material will become available in an online corpus search environment, similar to corpora already available through the National Centre for Human Language Technology (NCHLT) Corpus Portal.[3]

Possible aims of research that may follow the corpus building phase of the project are to identify the differences between the language use of L1 and L2 writers of the same language; and to describe the characteristics of learner language use at different levels of proficiency and/or at different stages of development (compare Timmis 2015, 125). With regard to contrastive analyses of this type, Johansson (2007, 2) points out that multilingual corpora provide a basis for empirical macro-linguistic studies comparing the realisation of, for instance, cohesion, speech acts, opening and closing of conversations etc. in different languages. Granger (2002, 22) asserts that, although traditionally only corpus linguists performed research based on learner corpora, this type of work is ideally suited for and in fact necessitates cooperative involvement of different disciplines. She writes (Granger 2002, 22): "With more and better learner corpora and truly interdisciplinary research teams there is no doubt that learner corpus research has the potential radically to improve knowledge about learner language and language learning." Applied research may include basic and advanced error analysis to inform reference sources such as learners' dictionaries (Granger and Paquot 2015) as well as digital language learning software (Granger 2015; Blanpain et al. 2017). Sophisticated descriptive and contrastive analysis of corpus material will also provide empirical support for much-needed curriculum transformation, aimed especially at improving the opportunities of linguistically at-risk students to succeed at university.

In order to bring together such a large and diverse corpus, thorough planning is necessary. The next section attempts to give an overview of the principles and criteria that were considered in the design of *SAMuLCAT*.

4. Design Considerations

4.1. General Considerations

With regard to learner corpora, Tono (2003, 801) cautions that "[t]here are quite a few projects in which not enough attention appears to have been paid to design considerations." He adds that "[i]f data is gathered in an opportunistic way without proper control and documentation of learner and task variables, the resulting corpus will be unlikely to be of much use."

Lozano and Mendikoetxea (2013, 90) also emphasise the need for "well-constructed large-scale learner corpora." According to these authors (2013, 89), such corpora should ideally be constructed within collaborative ventures by corpus linguists and SLA

[3] http://rmaservices.nwu.ac.za:8080/nchlt-whitelab/search/simple

researchers "to ensure that they are not simply opportunistic or ad hoc corpora." Lozano and Mendikoetxea (2013, 89), who have been instrumental in the design of the Spanish learner corpus *CEDEL2*, suggest that designers of learner corpora should start with general corpus design principles, such as the ten key design principles proposed by Sinclair (2004). Below, these principles are discussed with reference to *SAMuLCAT.* We reduced Sinclair's principles to eight for the sake of efficiency: "Content selection" and "topic" were subsumed under "external criteria"; and "balance" and "representativity" were combined, as they usually go hand in hand.

Principle 1: Use external criteria for content and topic selection

> The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function (Sinclair 2004).
>
> Any control of the subject matter in a corpus should be imposed by the use of external, and not internal, criteria (Sinclair 2004).

The content of *SAMuLCAT* will be determined by the disciplines within which the texts are produced, the genres and the topics of the assignments. For instance, an appropriate assignment to be written by students of the Natural Sciences could be an academic essay on the positive and negative effects of fracking; and an assignment topic for students in Health Care could be a patient history. Such assignments are not overtly aimed at eliciting specific collocations or certain types of errors.

Principle 2: Representativeness and balance

> The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components (Sinclair 2004).

McEnery, Xiao and Tono (2006, 73) agree that the notions of balance and representativeness are relative: A corpus should only be as representative as possible of the language variety under consideration. They conclude: "Corpus building is of necessity a marriage of perfection and pragmatism." However, the notion of representativeness is not always clear-cut, as many lecturers interpret the essay genre rather loosely, and texts uploaded as essays may not be aimed strictly at arguing for a certain position or expressing a personal opinion (compare Callies 2015, 45 with regard to the *ICLE*). Furthermore, although the ideal is to collect a balanced sample of genres produced in academic disciplines, the essay genre may be over-represented in learner corpora, as it is a popular pedagogic genre across disciplines at first-year level.

Principle 3: Contrast

> Only those components of corpora which have been designed to be independently contrastive should be contrasted (Sinclair 2004).

The metadata fields in *SAMuLCAT* (discussed in detail in the following section) have been designed to maximise comparison and contrast, for example texts written by L1 speakers versus texts written by L2 speakers; texts produced by students who obtained low marks for English in Grade 12, as opposed to texts produced by students who obtained high marks; and texts written by first-year students versus texts written by third-year students. On the other hand, since there is no metadata field to indicate collaboration between content subjects and Academic Literacy around assignments, it would not be meaningful to contrast material produced in content disciplines with material produced in Academic Literacy.

Principle 4: Structural criteria

> Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination (Sinclair 2004).

Granger (2012, 3235) emphasises the importance of compiling learner corpora according to strict design criteria pertaining to the learner and the task. "Multilingualism" is, for example, not a strict criterion. Timmis (2015, 119–120) sees a "multilingual" learner corpus as one in which "the data is gathered from speakers of several different languages" in contrast to a "monolingual" learner corpus consisting of "data gathered from speakers of one language." However, what is problematic about this definition is its ambiguous reference to the source and the target language of the learners. If Timmis refers to the L1s of the learners, the "multi-" of "multilingual" is not visible in the corpus, except in the metadata. Johansson's (2007, 9) definition focuses on the target languages of the corpora, and distinguishes between "translation corpora" and "comparable corpora": Translation corpora contain "original texts and their translations into one or more other languages," whereas comparable corpora contain "original texts in two or more languages matched by criteria such as genre, time of publication, etc." Schmidt and Wörner (2012, xi), in the introduction to their study on multilingual corpora, define the term "multilingual" flexibly:

> In order to qualify as multilingual, a corpus thus need not necessarily contain texts in more than one language—monolingual data can also tell us something about multilingualism if the texts are produced by a multilingual speaker, if they arise from a multilingual communicative setting or if they are put in relation to the multilingual society in which they are observed. In that sense, multilinguality is not simply an intrinsic property of the language data contained in a corpus, but rather a consequence of how a corpus is designed, documented and used.

The envisaged corpus embodies and embraces "multilinguality" as a design principle, in that the texts are produced by speakers of different first languages, who may be proficient in several other languages, whose language production may have been influenced by several languages in a multilingual communicative setting, and who may produce texts in more than one language.

Principle 5: Annotation

> Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications (Sinclair 2004).

In order for corpora such as *SAMuLCAT* to be fully utilised by modern corpus linguistic software, the corpora will be automatically annotated on two linguistic levels, namely part-of-speech and lemmas. For English content, each token in the corpus will be annotated by a state-of-the-art lemmatiser (Müller et al. 2015), after which the tokens are assigned a part-of-speech tag by the NLP4J open-source part-of-speech (POS) tagger (Choi 2016). Content in the other South African languages will be annotated by using the language-specific NCHLT lemmatisers (Eiselen and Puttkammer 2014), and improved NCHLT POS taggers (Puttkammer et al. 2018). Although these automatic processes are not perfect, the addition of annotations allows for much richer corpus analytics by allowing compound search heuristics where any combination of token, lemma, and part-of-speech can be used in the analysis of the corpus. However, it should be kept in mind that automatic POS annotation in the case of a learner corpus is only a rough first indication of where the target forms are potentially located in the corpus (Van Rooy 2015, 86); annotation errors might occur as a result of "the idiosyncratic properties of the learner language" (Van Rooy 2015, 88).

Principle 6: Sample size

> Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size (Sinclair 2004).

For *SAMuLCAT*, the text has to be a full genre or part-genre (for example an essay, a speech or a paragraph with a particular rhetorical function, and not an utterance or a few sentences). The minimum length will be 250 words.

Principle 7: Documentation

> The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken (Sinclair 2004).

This article may be regarded as part of the documentation that delineates the content and arguments in justification of the choices made. With the release of the corpus, further documentation relating to the composition, content, format, and annotations will be made available. This documentation will be distributed with the associated corpus, as well as in further scientific publications regarding the corpus.

Principle 8: Homogeneity

> A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts [unusual texts which stand out as radically different from the others—authors] should be avoided (Sinclair 2004).

The notion of "homogeneity" may not be easily attainable with regard to a multi-L1, multi-L2, multimodal, multi-level and multi-genre corpus. However, the corpus managers will maintain the right to discard any texts that may jeopardise the principle of homogeneity.

Gilquin (2015, 10) adds the notion of "naturalness" (authenticity) as a criterion that is specific to learner corpora. In stating that there are many variables that may affect naturalness, she links her assertion to Granger's (2012, 3235) "degrees of naturalness." A text produced in response to a prompt in a study guide may be natural in that it is the authentic linguistic output of a foreign- or second-language learner, but the message may not be authentic. Naturalness may also be affected by students' reliance on various forms of assistance, for instance the structure, content and wording of scholarly articles, textbooks and dictionaries, lecturer feedback, writing centre assistance, or even editing. Another issue relevant to naturalness is whether texts written by a schooled adult user in an "indigenised variety" of a language counts as a learner text. Gilquin (2015, 12) argues that a corpus produced by adolescent undergraduate university students who are L2 speakers of the target language "may be justified," but questions whether texts produced by adult speakers of the same target language would qualify. This reservation casts doubt on whether drafts of theses and dissertations by L2 writers qualify as learner texts.

The next section focuses on specific design considerations—particularly those that will determine the structure of *SAMuLCAT*.

4.2. Specific Design Considerations

Tono (2003, 800) proposes three major design considerations or categories in relation to which designers of especially learner corpora should make decisions (each with a number of sub-categories), viz. language-, task- and learner-related criteria. Gilquin (2015) omits language-related criteria, and distinguishes environment, task and learner variables. Target language, for instance, is an environmental criterion (Gilquin 2015, 16). For *SAMuLCAT* we shall regard the target language as a task-related variable, as the target language is determined by the task at hand. In the remainder of this section we shall use Tono's typology as a starting point, while justifying our own classification as the discussion unfolds.

According to Tono (2003), *language-related* variables include medium, genre, style and topic. However, medium/mode and genre are usually inextricably bound up with the task at hand (Gilquin 2015, 16), and style is determined by genre. We shall therefore categorise all of the above as task-related variables, and discard the category "language-related."

Tono's (2003) *learner-related* criteria include internal-cognitive (age/cognitive style), L1 background, L2 environment (ESL/EFL) and L2 language proficiency (e.g. determined by test score). For the purpose of *SAMuLCAT*, L1 background and L2 language proficiency will be treated as learner-related variables, while L2 environment/target language will be regarded as a task-related variable.

The measurement of L2 proficiency is handled differently by different corpus designers. Granger (1998, 9) reports that for the *ICLE*, L2 proficiency is determined by variables such as age and the number of years the student has studied English at university. Other corpora require a score on an independent, standardised proficiency test; for example, *CEDEL2* requires the University of Wisconsin Placement Test (Lozano and Mendikoetxea 2013, 74). Although it is conceded that a single independent measure is the most reliable way to determine proficiency in the target language, it would not be attainable for a multilingual South African corpus, as equivalent proficiency tests have not yet been developed for all the South African indigenous languages. At least initially, we shall rely on the learner's self-reported score for the particular language at NSC (National Senior Certificate) or equivalent level, provided that the target language had been passed as an NSC subject or a subject recognised by an equivalent certification body.

Tono's (2003) *task-related criteria* include data collection (cross-sectional/longitudinal), elicitation (spontaneous/prepared) and time limitation (fixed/free/homework). We decided to include these, and based on the arguments presented above, to add target language, text type/genre, version and mode. A number of the task-related variables selected for *SAMuLCAT* will be discussed in some detail below, viz. target language, version and mode.

Although early learner corpora were predominantly focused on English as the target language, a number of projects that have in recent years been launched focus on other (single) target languages, for instance German (*Falko, Fehlerannotiertes Lernerkorpus*), French (*FLLOC, French Learner Language Oral Corpora*) and Spanish (*CEDEL2, Corpus Escrito del Español L2*). An even more recent addition to the typological category of "learner corpus" is the multilingual learner corpus (Gilquin 2015, 13). Examples are the *MiLC Corpus*, which contains learner data in Catalan, English, French and Spanish as target languages, and the *USP Multilingual Learner Corpus*, which has English, German, Italian and Spanish as target languages.

Two metadata fields in the corpus focus on the distinction between cross-sectional and longitudinal data collection, viz. "version" and "current university level." "Cross-sectional" concerns data-gathering at a particular moment in students' developmental trajectories, whereas "longitudinal" refers to a particular student's development over time. The majority of learner corpora are cross-sectional. An example of a longitudinal corpus is the *Longitudinal Database of Learner English* (*LONGDALE*), which follows the development of the same learners over a minimum period of three years, with at least one data collection per year (Gilquin 2015, 14). There are not many pure longitudinal learner corpora, since students may drop out during the course of the longitudinal data collection; and thus, some compilers revert to pseudo-longitudinal corpora (Gass and Selinker 2008, 56–57). Such corpora are gathered from users with different proficiency levels (for example first year, second year, etc.) at a specific point in time. Corpora may also be hybrid in that they contain cross-sectional, pseudo-longitudinal and pure longitudinal data. The first version of *SAMuLCAT* will be primarily cross-sectional, as the bulk of the assignments uploaded over the collection period will have been written by different cohorts of first-year students. Some pseudo-longitudinal data may be added in the format of texts composed by students at higher levels of maturity (2nd year to doctoral level), and pure longitudinal data will be included where revisions based on lecturer feedback (different versions) are sequentially uploaded.

The notion of "mode" will be included with task-related variables, as the choice of a particular mode is usually determined by the academic discipline, the topic, the genre and the outcomes of the course or module. Motivating students to produce work in modes other than written text has been given momentum by the multimodal turn in teaching, learning and research. Additionally, it is becoming more and more popular to collect data on oral and audio-visual presentations by means of video-recordings, which are then transcribed. Apart from speech, the transcription may involve representations of embodied modes, such as gestures, posture, gaze, etc. If multimodal corpora are distributed with their video files, "it is possible and often desirable) to align the text transcript with the sound/video so that the two can be examined and queried simultaneously" (Gilquin 2015, 21–22).

In following Granger (2012), a main distinction will be made between learner-related and task-related information in *SAMuLCAT*, with the following sub-categories and variable values:

Learner-related information that forms part of the *SAMuLCAT* metadata includes the following: unique identifier; institution; degree programme; level (first year–PhD); age; home language (in addition to the other ten official South African languages we decided to include English, as it was argued that all first-year students are to some extent learners of the academic English of scientific disciplines); language(s) passed at home language level for the NSC; father's first language; mother's first language; NSC symbol for English (Home Language or First Additional Language); medium of instruction at school; and type of school attended (city/town, rural, home).

The following task-related metatdata categories (which include Gilquin's [2015] "environmental variables") are included in *SAMuLCAT*: date of the assignment; institution at which the assignment was submitted; subject/module for which the assignment was produced (Academic Literacy or a content module); disciplinary content (e.g. Education, Engineering, Law); target language (a choice from all eleven official languages); mode (written, spoken or multimodal); collaboration (individual or group assignment); version (1, 2, 3, etc.); intervention or support (no intervention, lecturer, writing centre, peer, CALL software, library, dictionary, spell checker); location (home, university residence or class); time limitation; and word count.

The next section focuses on the development of the platform and collection procedures for *SAMuLCAT*.

5. Data Collection and Storage

Institutions usually expect that projects involving human respondents should be ethically cleared by the relevant authorities, e.g. an Ethics Committee. In the case of the present project, two South African universities had approved the data collection, data analysis, reporting, and storage protocol of the project at the time this article was submitted for publication. The protocol specifies that learner data will be collected by recruiting learners from amongst the students to which the compilers of the corpus have access. Although it is expected that entire populations should contribute texts, the selection will eventually be based on those who give consent for their texts to be used as data, and the texts that fulfil the criteria established for the corpus design.

Although typewritten texts can be scanned and converted through optical character recognition, or uploaded in pdf format, it is nowadays customary for learner-generated texts to be uploaded to a particular platform or via a specific link on the online interface of their institution in word processed format or text only. If the texts are uploaded via the learning management system of an institution, they can be downloaded in bulk and transferred to a centralised data management system, which is responsible for all of the corpus content, as well as for collecting all of the required metadata. The metadata can either be integrated in the content directly, or included in a database which is linked to the relevant content files. This integration of the metadata allows researchers to identify relevant content based on any number of variables, in the form of metadata, and then extract the part of the corpus that corresponds to these criteria (Gilquin 2015, 18). The latter process is, for instance, used by the *ICLE*.

The data management system facilitating the collection and integration of metadata consists of three main components. Firstly, there is a website where learners provide the metadata as described in the previous section. A second website is used by lecturers to submit the metadata for each task that is submitted to the data management system. The final component in the data management system is an integration module, which combines all of the submitted metadata and the student writing assignments. Figure 1 shows the high-level architecture of the data management system.

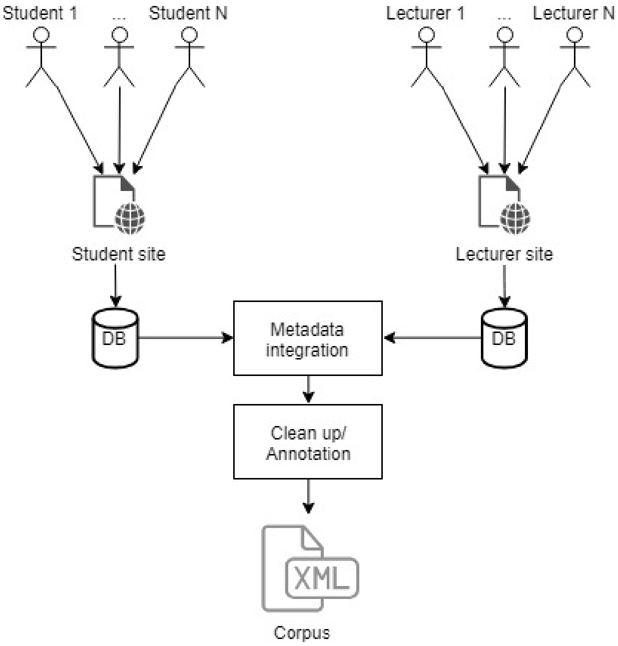

Figure 1: Collection architecture overview

The following sections provide a brief overview of each of these components, some of the design decisions that were taken during development, and the interaction between each of the components.

5.1. Learner-Related Metadata Service

The first site developed for the creation of the corpus is a learner-related metadata website, which is a very basic interface for collecting biographic information from students, as described in 4.2. The primary concern in developing the site was to make the interface as user-friendly as possible, while limiting the number of errors that may occur. Furthermore, no identifiable personal information, such as names, birth dates and student numbers are stored, while still allowing the linking of specific learner-related metadata with specific learner task submissions. Lastly, it is imperative that the process of entering metadata by the students is as efficient as possible, in other words that it is not too time-consuming or complex, while also limiting security risks for participating institutions. For these reasons, the metadata collection service was not integrated into the universities' official record systems and teaching and learning platforms.

5.2. Task-Related Metadata Service

The second website developed for *SAMuLCAT* is the site to collect task-related metadata, as described in 4.2. This site is only made available to the lecturers, tutors and assistants who are responsible for the administration of the modules within which task materials are collected. The interface for the task-related collection service follows design principles similar to those for the learner-related information in order to minimise errors, while also being as comprehensive and as flexible as possible in recording the various types of metadata required. For each task submitted to the system, a lecturer or assistant completes the set of metadata associated with the specific task, and uploads this form together with the bulk set of assignments submitted by the learners to the *SAMuLCAT* platform. The task-related metadata, along with the assignments, are then securely stored in an encrypted database that is used in the corpus integration module.

5.3. Corpus Integration Module

The final stage in the corpus creation software pipeline is the corpus integration and creation module. The main purpose of this module is linking the task-related and learner-related metadata, as well as the individually submitted tasks of each learner, into a coherent, standardised corpus. For each task submitted to the task-related database, each of the individual assignments associated with the task is linked to the specific metadata instance submitted by a specific learner. This means that each assignment includes the specific metadata of the learner who submitted the assignment. This further implies that on release of the corpus it will be possible to identify all assignments that satisfy one or more of the metadata criteria. It will, for instance, be possible to identify assignments based both on task- and learner-related metadata simultaneously, for instance all assignments created by learners with isiZulu as their home language (learner-related), within the essay genre, submitted in English in the field of Education (task-related).

The final step in the corpus creation process is the automatic annotation of the data on four levels, namely sentence, token, part of speech, and lemma, using state-of-the-art English modules developed by Choi (2016) and Müller et al. (2015). Annotation is technically "the assignment of a category to a segment of the corpus" (Lüdeling and Hirschmann 2015, 136). Annotations make it possible to retrieve categories from the data that would otherwise be cumbersome or impossible to retrieve (Van Rooy 2015,

83) and provide additional information that may be valuable resources for more fine-grained linguistic analyses of the source data. Annotations are separable from the raw data, making it possible to retrieve the original corpus content.

The output of the data management system is an XML file that complies with the widely-accepted and standardised TEI P5 format (Burnard 2010). This format allows for relatively simple extension and enrichment of the corpus with additional metadata and linguistic annotations; most importantly, the TEI format can easily be integrated and reused by various freely available corpus analysis tools.

It is also important to note that although the primary aim of the metadata services and corpus integration module is to process English data, the modules have been developed with the expectation that they will be extended to other languages. The services and module are designed to be immediately extensible to the other South African languages, allowing for automatic linguistic annotation using modules developed by Eiselen and Puttkammer (2014). This means that the same system and process can be used by other researchers and research groups to create similar corpora in any of the South African languages.

6. Limitations

The first phase of the project focused on designing the corpus, with specific reference to metadata categories; planning and setting up the data management system to facilitate the collection of the learner corpus texts and integration of the metadata; and uploading a minimum of 4 000 texts (approximately 2 million words) with all the required metadata. The first two goals have been achieved, and approximately 1 500 academic learner texts have been uploaded.

Limitations are that the uploaded texts, and those that will still be uploaded during the first phase of the project, comprise mostly first-version essays written by first-year students enrolled for subject-specific modules in Academic Literacy. Although these texts have a broad subject-specific focus, for example topics related to subjects in the Natural Sciences and Economic and Management Sciences, it would be necessary to also include texts written in academic disciplines, also at more advanced levels than first year; texts written in other genres than argumentative essays; texts produced in modes other than written text; and texts produced in languages other than English and Afrikaans—particularly the indigenous African languages.

7. Next Steps

The immediate next step, which may still take place within the current phase of the project, will include a broadening of the scope of the corpus in terms of specificity, genre, level, focal language and mode. Concurrently, additional layers of information in the format of annotations will be added to make the corpus more useful for research and applications.

In addition to forms of annotation that are aligned with typical levels of linguistic analysis, such as POS tagging and syntactic parsing, a number of problem-oriented annotation systems for learner corpora can be distinguished, of which error tagging is one of the most valuable systems. Error annotation systems are based on error taxonomies which contain categories for error classification (Díaz-Negrillo and Fernández-Domínguez 2006, 92). The construction of an error annotation system involves the design of a taxonomy of errors alongside its related tags, which have to be inserted in the learner corpus. Although some attempts have been made at automating error tagging, the general practice is to use an editor that facilitates computer-assisted manual insertion of tags. One of the editors that have recently been developed contains tag-associated error categories that are arranged on a menu-driven interface. The annotator then selects and inserts tags in the text, and decides on the nature of each of the identified errors (Díaz-Negrillo and Fernández-Domínguez 2006, 86). In addition to selecting or designing an error taxonomy with a related tagset, corpus designers should decide whether they also wish to insert corrections or reconstructions of the errors, omissions and redundancies.

Before embarking on the design of an error tagging taxonomy, and attempting to develop a tagset, the designers of *SAMuLCAT* need to answer the following questions as steps in the decision-making process:

1. *Should the taxonomy serve multiple purposes, or is there a pertinent research question to be answered?*
   This question is important in light of the fact that one of the immediate applications of the corpus is the development of language-specific computerised CALL tools that automatically detect frequently occurring language-specific errors, and provide suggestions for improvement.
2. *Is there an existing computer-aided error analysis program (CEA) that is suitable for the annotation of SAMuLCAT in light of the answer to 1?*
   Regarding written text, this question can only be answered after having studied the dimensions that are generally included in existing error taxonomies (Tono 2003; Díaz-Negrillo and Fernández-Domínguez 2006), their structure and the specific error tags used (Díaz-Negrillo and Fernández-Domínguez 2006, 92ff). Existing CEA programs include the *Louvain*, *Free-Text*, *CLC*, *Falko* and *NICT JLE* (National Institute of Information and Communications Technology Japanese Learner of English Corpus) taxonomies. It seems that only the *Louvain* error tagging system has been commercialised.

3. *If there is no existing CEA that will answer the research questions asked by the SAMuLCAT design team, how will a project-specific taxonomy be structured and what will the main categories be?*
   One way of addressing research questions related to computerised feedback on learner assignments is to use error lists compiled by lecturers with many years of experience in assessing academic literacy assignments and exams. The categories identified inductively may then be compared to tagsets used for other corpora of learner texts, for example as contained in the "Error Tagging Manual" for the *ICLE* (Version 1.2) (Dagneaux et al. 2008). Ideally the tagset should be sufficiently general to allow its use on texts in all the languages represented in the corpus (Tono 2003).

Regarding modalities other than written text, for instance audio- and video-recorded learner data, commercially available software for transcription and annotation will have to be evaluated for their utility and efficiency, for example ELAN and EXMARaLDA for multimodal data (Ballier and Martin 2015, 123)

Technical challenges that will have to be addressed before annotation can commence include the integration of the tags with the corpus. According to Smith, Hoffmann and Rayson (2008, 167) manual annotations are often not integrated with the actual corpus. They are typically treated in separate spreadsheet files and other database structures. Thus it seems advisable that, similar to the *Free-Text* System, the *CLC* and the *NICT JLE*, *SAMuLCAT* should ideally make use of XML tags.

Acknowledgements

This research was made possible with support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Technology of South Africa as part of the South African Research Infrastructure Roadmap (SARIR).

We would like to thank Prof Tobie van Dyk for his management and co-ordination of this project, which falls under the ICELDA node within SADiLaR; Dr Elsa Meihuizen for her valuable input on multilingual corpora; and Dr Zanette Meintjes for her contribution on learner corpora with Afrikaans as target language. Both colleagues also provided very useful feedback on the metadata fields designed for the corpus.

References

Alsop, S., and H. Nesi. 2009. "Issues in the Development of the British Academic Written English (BAWE) Corpus." *Corpora* 4 (1): 71–83. https://doi.org/10.3366/E1749503209000227.

Ballier, N., and P. Martin. 2015. "Speech Annotation of Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 107–134. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.006.

Blanpain, K., A. Laffut, S. Verlinde and L. De Wacther. 2017. "De ILT AcademicWritingAssistant. Een tool voor het schrijven van academische teksten in het Engels." Accessed August 6, 2017. http://www.nut-talen.eu/docentendag2017/NUT-Docentendag%20-%20KU%20Leuven.pdf.

Burnard, L. 2010. "TEI P5: Guidelines for Electronic Text Encoding and Interchange P5, Version 1.6." TEI Consortium. Accessed February 12, 2018. http://www. tei-c. org/Guidelines/P5/.

Callies, M. 2015. "Learner Corpus Methodology." In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 35–55. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003.

Centre for English Corpus Linguistics. n.d. "Learner Corpora around the World." Louvain-la-Neuve: Université catholique de Louvain. Accessed July 16, 2017. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Choi, J. D. 2016. "Dynamic Feature Induction: The Last Gist to the State-of-the-Art." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 271–281. San Diego, June 12–17, 2016. https://doi.org/10.18653/v1/N16-1031.

Dagneaux, E., S. Denness, S. Granger, F. Meunier, J. Neff, and J. Thewissen. 2008. "Error Tagging Manual Version 1.3." http://hdl.handle.net/2078.1/75586.

Dahlmeier, D., H. T. Ng, and S. M. Wu. 2013. "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English." In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31. Atlanta, June 13, 2013. http://www.aclweb.org/anthology/W13-1703.

Department of Higher Education and Training. 2017. *Draft Revised Language Policy on Higher Education*. Pretoria: Department of Higher Education and Training.

Díaz-Negrillo, A., and J. Fernández-Domínguez. 2006. "Error Tagging Systems for Learner Corpora." *RESLA* 19: 83–102.

Eiselen, R., and M. J. Puttkammer. 2014. "Developing Text Resources for Ten South African Languages." In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 3698–3703. Reykjavik, May 26–31, 2014. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf.

Gass, S. M., and L. Selinker. 2008. *Second Language Acquisition: An Introductory Course*. 3rd ed. New York: Routledge.

Gilquin, G. 2015. "From Design to Collection of Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 9–34. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002.

Granger, S. 1998. "The Computer Learner Corpus: A Versatile New Source of Data for SLA Research." In *Learner English on Computer*, edited by S. Granger, 3–18. London: Longman.

Granger, S. 2002. "A Bird's-Eye View of Learner Corpus Research." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, edited by S. Granger, J. Hung and S. Petch-Tyson, 3–36. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.6.04gra.

Granger, S. 2012. "Learner Corpora." In *The Encyclopedia of Applied Linguistics*, edited by C. A. Chapelle, 3235–3242. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0669.

Granger, S. 2015. "The Contribution of Learner Corpora to Reference and Instructional Materials Design." In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 486–510. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022.

Granger, S., E. Dagneaux and, F. Meunier. 2002. *The International Corpus of Learner English. Version 1.1. Handbook and CD-ROM.* Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., and M. Paquot. 2015. "Electronic Lexicography Goes Local: Design and Structures of a Needs-Driven Online Academic Writing Aid." *Lexicographica* 31 (1): 118–141. https://doi.org/10.1515/lexi-2015-0007.

Huang, L-S. 2017. "Taking Stock of Corpus-Based Instruction in Teaching English as an International Language." *RELC Journal*. https://doi.org/10.1177%2F0033688217698294.

Johansson, S. 2007. *Seeing through Multilingual Corpora: On the Use of Multilingual Corpora in Contrastive Studies*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.26.

Lozano, C., and A. Mendikoetxea. 2013. "Learner Corpora and Second Language Acquisition." In *Automatic Treatment and Analysis of Learner Corpus Data*, edited by A. Díaz-Negrillo, N. Bailler and P. Thompson, 65–100. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.59.06loz.

Lüdeling, A., and H. Hirschmann. 2015. "Error Annotation Systems." In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 135–157. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.007.

McEnery, T., R. Xiao, and R. Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book.* London: Routledge.

Müller, T., R. Cotterell, A. Fraser, and H. Schütze, 2015. "Joint Lemmatization and Morphological Tagging with LEMMING." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2268–2274. Lisbon, Portugal, September 17–21, 2015. https://doi.org/10.18653/v1/D15-1272.

Puttkammer, M. J., E. R. Eiselen, J. Hocking, and F. J. Koen. 2018. "NLP Web Services for Resource-Scarce Languages." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, pp. 43–49. Melbourne, Australia, July 15–20, 2018. http://aclweb.org/anthology/P18-4008.

Richards, J. C., and T. S. Rodgers. 2014. *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Römer, U. 2012. "Using General and Specialised Corpora in English Language Teaching: Past, Present and Future." In *Corpus-Based Approaches to English Language Teaching,* edited by M. C. Campoy-Cubillo, B. Bellés-Fortuño, and M. L. Gea-Valor, 18–38. New York: Continuum.

Roux, J. C. 2016. "South African Centre for Digital Language Resources." In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2467–2470. Portorož, Slovenia, May 23–28, 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/337_Paper.pdf.

Schmidt, T., and K. Wörner. 2012. Introduction to *Multilingual Corpora and Multilingual Corpus Analysis*, edited by T. Schmidt and K. Wörner, xi–xiii. Amsterdam: John Benjamins. https://doi.org/10.1075/hsm.14.01intro.

Sinclair, J. 2004. "Corpus and Text—Basic Principles." In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by M. Wynne. Oxford: Oxbow. http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm.

Smith, N., S. Hoffmann, and P. Rayson. 2008. "Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations." *Literary and Linguistic Computing* 23 (2): 163–180. https://doi.org/10.1093/llc/fqn004.

Timmis, I. 2015. *Corpus Linguistics for ELT.* London: Routledge. https://doi.org/10.4324/9781315715537.

Tono, Y. 2003. "Learner Corpora: Design, Development and Applications." In *Proceedings of the Corpus Linguistics 2003 Conference*, edited by D. Archer, P. Rayson, A. Wilson and A. M. McEnery, 800–809. http://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdfhttp://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdf.

Van Rooy, B. 2015. "Annotating Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, S. Gilquin and F. Meunier, 80–105. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.005.