GHENT UNIVERSITY

# Course Specifications

From the academic year 2024-2025 up to and including the academic year

## Bantu Corpus Linguistics and Lexicography (A005117)

**Course size** *(nominal values; actual values may depend on programme)*

**Credits 5.0** **Study time 150 h**

**Course offerings in academic year 2025-2026**

A (semester 1) English Gent

**Lecturers in academic year 2025-2026**

de Schryver, Gilles-Maurice LW21 lecturer-in-charge

| Offered in the following programmes in 2025-2026 | crdts | offering |
|---|---|---|
| Master of Science in Teaching in Languages(main subject African Languages and Cultures) | 5 | A |
| Master of Arts in African Studies | 5 | A |
| Master of Arts in Advanced Studies in Linguistics (main subject Linguistics in a Comparative Perspective) | 5 | A |
| Exchange Programme African Languages and Cultures | 5 | A |

**Teaching languages**

English

**Keywords**

Bantu languages, corpus linguistics, lexicography, data-driven language technology

**Position of the course**

This is an advanced course in which corpus linguistics for the Bantu languages as well as data-driven lexicography is introduced. Students acquire an in-depth knowledge of the methodologies, the tools, as well as the strengths and limitations of the analytical apparatus, and are invited to put the acquired knowledge into practice.

**Contents**

Corpus linguistics is a booming field, well covered in both basic textbooks (e.g. Biber *et al.* 1998, Kennedy 1998, McEnery & Wilson 2001) as well as in more advanced studies (e.g. McEnery *et al.* 2006, Renouf & Kehoe 2009, McEnery & Hardie 2012), which has largely been used to investigate the world's major languages (e.g. Sinclair 1991, Meyer 2002, O'Keeffe & McCarthy 2010). However, even in collections which aim at sampling the world's languages, the result deals largely with European languages (e.g. Wilson *et al.* 2006). The same can, mutatis mutandis, be said about data-driven lexicography. In this advanced course this status quo is challenged by focusing on corpus linguistics and lexicography for the Bantu language family. Given the state of the discipline, results from the wider field of African language technology (De Pauw *et al.* 2006-16) are also brought in. Each class consists of two parts. In the first a topic form the list below is considered, and in the second the newly acquired knowledge is immediately put into practice on the computer.

- Class 1 – **Bantu CORPORA: What, how and use?** (de Schryver & Prinsloo 2000a, de Schryver 2002)
- Class 2 – **SOFTWARE for Bantu corpus linguistics and data-driven lexicography** (Scott 1996-2016, Joffe 2002-16, de Schryver & De Pauw 2007)
- Class 3 – **Bantu corpus APPLICATIONS: Fundamental research, teaching and language learning** (Prinsloo & de Schryver 2001)
- Class 4 – **Bantu SPELLCHECKERS: non-word error detection** (Prinsloo & de Schryver 2003)
- Class 5 – **Bantu corpus TERMINOGRAPHY** (Taljard & de Schryver 2002)

- Class 6 – **Bantu corpus-based TRANSLATION studies** (Gauton *et al.* 2003)
- Class 7 – **Bantu corpus LEXICOGRAPHY 1: Basic aspects** (de Schryver & Prinsloo 2000b, 2000c)
- Class 8 – **Bantu corpus LEXICOGRAPHY 2: Advanced aspects** (de Schryver & Joffe 2004, de Schryver *et al.* 2006)
- Class 9 – **Bantu corpus LINGUISTICS 1: Synchronic aspects** (de Schryver & Nabirye 2010)
- Class 10 – **Bantu corpus LINGUISTICS 2: Diachronic aspects** (de Schryver & Gauton 2002)
- Class 11 - **Bantu corpus LINGUISTICS 3: Strengths** (Kawalya *et al.* 2014)
- Class 12 – **Bantu corpus LINGUISTICS 4: Limitations** (Bostoen & de Schryver 2015)

Through state-of-the-art literature students are able to familiarize themselves with the early stages, development, and current use of data-driven methods in Bantu linguistics and lexicography. The differences with other approaches to data collection, data analysis and data synthesis (including questionnaires, stimuli, grammaticality judgement tests, introspection and intuition) are also given due attention.

Knowledge of Bantu languages or other African languages is an advantage, but not a prerequisite to follow this course. Each week students are asked to read one or more journal articles or book chapters as preparation for the lesson. The contents of these articles and chapters are discussed during class and put in a broader perspective.

**Initial competences**

'Bachelor in African Languages and Cultures', 'Bachelor in Linguistics', or equivalent.

**Final competences**

An advanced knowledge and understanding of Bantu corpus linguistics and lexicography.

**Conditions for credit contract**

Access to this course unit via a credit contract is unrestricted: the student takes into consideration the conditions mentioned in 'Starting Competences'

**Conditions for exam contract**

This course unit cannot be taken via an exam contract

**Teaching methods**

Seminar

**Study material**

Type: Reader

Name: Scholarly articles and book chapters are provided on the Ufora course site.
Indicative price: Free or paid by faculty
Optional: no
Language : English
Available on Ufora : Yes
Online Available : Yes
Available in the Library : Yes

**References**


- Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge Approaches to Linguistics). Cambridge: Cambridge University Press.
- Bostoen, K. & de Schryver, G.-M. 2015. 'Linguistic innovation, political centralization and economic integration in the Kongo kingdom: Reconstructing the spread of prefix reduction'. *Diachronica* 32,2: 139–85 + 13 pages of supplementary material online.
- De Pauw, G., de Schryver, G.-M. & Wagacha, P.W. 2006-16. *AfLaT - African Language Technology*. Retrieved from http://aflat.org/.
- de Schryver, G.-M. 2002. 'Web for/as corpus: A perspective for the African languages'. *Nordic Journal of African Studies* 11,2: 266–82.
- de Schryver, G.-M. & De Pauw, G. 2007. 'Dictionary writing system (DWS) + corpus query package (CQP): The case of TshwaneLex'. *Lexikos* 17: 226–46.

- de Schryver, G.-M. & Gauton, R. 2002. 'The Zulu locative prefix ku- revisited: A corpus-based approach'. *Southern African Linguistics and Applied Language Studies* 20,4: 201–20.
- de Schryver, G.-M. & Joffe, D. 2004. 'On How Electronic Dictionaries are Really Used'. In G. Williams & S. Vessier (eds), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, pp. 187–96.
- de Schryver, G.-M., Joffe, D., Joffe, P. & Hillewaert, S. 2006. 'Do Dictionary Users Really Look Up Frequent Words? – On the Overestimation of the Value of Corpus-based Lexicography'. *Lexikos* 16: 67–83.
- de Schryver, G.-M. & Nabirye, M. 2010. 'A quantitative analysis of the morphology, morphophonology and semantic import of the Lusoga noun'. *Africana Linguistica* 16: 97–153.
- de Schryver, G.-M. & Prinsloo, D.J. 2000a. 'The compilation of electronic corpora, with special reference to the African languages'. *Southern African Linguistics and Applied Language Studies* 18,1-4: 89–106.
- de Schryver, G.-M. & Prinsloo, D.J. 2000b. 'Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure'. *South African Journal of African Languages* 20,4: 291–309.
- de Schryver, G.-M. & Prinsloo, D.J. 2000c. 'Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure'. *South African Journal of African Languages* 20,4: 310–30.
- Gauton, R., Taljard, E. & de Schryver, G.-M. 2003. 'Towards Strategies for Translating Terminology into all South African Languages: A Corpus-based Approach'. In G.-M. de Schryver (ed.), *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*. Pretoria: (SF)2 Press, pp. 81–8.
- Joffe, D. 2002-16. *tlCorpus - Concordance Software*. Retrieved from http://tshwanedje.com/corpus/.
- Kawalya, D., Bostoen, K. & de Schryver, G.-M. 2014. 'Diachronic semantics of the modal verb -sóból- in Luganda: A corpus-driven approach'. *International Journal of Corpus Linguistics* 19,1: 60–93.
- Kennedy, G. 1998. *An Introduction to Corpus Linguistics* (Studies in Language & Linguistics). Harlow, Essex: Addison Wesley Longman.
- McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- McEnery, T. & Wilson, A. 2001. *Corpus Linguistics: An Introduction. 2nd Edition* (Edinburgh Textbooks in Empirical Linguistics). Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.
- Meyer, C.F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- O'Keeffe, A. & McCarthy, M. (eds). 2010. *The Routledge Handbook of Corpus Linguistics*. Milton Park, Abingdon: Routledge.
- Prinsloo, D.J. & de Schryver, G.-M. 2001. 'Corpus applications for the African languages, with special reference to research, teaching, learning and software'. *Southern African Linguistics and Applied Language Studies* 19,1-2: 111–31.
- Prinsloo, D.J. & de Schryver, G.-M. 2003. 'Non-word error detection in current South African spellcheckers'. *Southern African Linguistics and Applied Language Studies* 21,4: 307–26.
- Renouf, A. & Kehoe, A. (eds). 2009. *Corpus Linguistics: Refinements and Reassessments*. Amsterdam: Rodopi.
- Scott, M. 1996-2016. *WordSmith Tools*. Retrieved from http://www.lexically.net/wordsmith/.
- Sinclair, J. 1991. *Corpus, Concordance, Collocation* (Describing English Language). Oxford: Oxford University Press.
- Taljard, E. & de Schryver, G.-M. 2002. 'Semi-automatic term extraction for the African languages, with special reference to Northern Sotho'. *Lexikos* 12: 44–74.
- Wilson, A., Archer, D. & Rayson, P. (eds). 2006. *Corpus Linguistics around the World*. Amsterdam: Rodopi.


**Course content-related study coaching**

Feedback during and after class, via e-mail, and through the Ufora course site

**Assessment moments**

continuous assessment

**Examination methods in case of periodic assessment during the first examination period**

Assignment

**Examination methods in case of periodic assessment during the second examination period**

Assignment

**Examination methods in case of permanent assessment**

Participation, Assignment

**Possibilities of retake in case of permanent assessment**

not applicable

**Calculation of the examination mark**

50% permanent evaluation (class participation, advance reading of provided literature)
50% research paper in which the in-depth knowledge of Bantu corpus linguistics and data-driven lexicography that was acquired in this course is applied to one or more Bantu languages of the student's choice (minimum 6000 words, maximum 10,000 words, excluding references and addenda)

**Facilities for Working Students**

None: hands-on participation is required.