Bibliography

Aarts, J, and Meijs, W, (eds.). 1990. *Theory and Practice in Corpus Linguistics*. Vol. 4, *Language and Computers: studies in practical linguistics*. Amsterdam: Rodopi.

Aitchison, J. 1992. Teach yourself linguistics. London. Hodder & Stoughton.

Algeo, J. 1988. British and American grammatical differences. *International Journal of Lexicography* 1 (1):1-31.

Alidou, H. 2004. Medium of instruction in post-colonial Africa. In *Medium of Instruction Policies: Which agenda? Whose Agenda?*, edited by J. W. Tollefson and A. B. M. Tsui. Mahwah: Lawrence Erlbaum Associates.

Allwood, J. 1998. Some frequency based differences between spoken and written Swedish. Paper read at The xvi:th Scandinavian conference of Linguistics, at University of Turku.

Al-Sulaiti, L. 2004. Designing and Developing a Corpus of Contemporary Arabic. MSc. thesis, School of Computing, University of Leeds, Leeds.

Andersson, L-G, and Janson, T. 1997. *Languages in Botswana; language ecology in Southern Africa*. Gaborone: Longman Botswana.

Archbell, J. 1837. *A grammar of the Bechuana language*. Cape of Good Hope: Meurant and Godlonton.

Arthur, J. 1997. 'There must be something undiscovered which prevents us from doing our work well': Botswana primary teachers' views on educational language policy. *Language and education* 11 (4):225-241.

Arua, A.E, and Magocha, K. 2002. Patterns of language use and language preference of some children and their parents in Botswana. *Journal of multilingual and multilingual development* 23 (6):449-461.

Aston, G. 2001. Text categories and corpus users: A response to David Lee. *Language Learning and Technology* 5 (3):73-76.

Aston, G, and Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. T. McEnery and A. Wilson (Eds), *Edinburgh Textbooks in Empirical Linguistics*. Edinburgh: Edinburgh University Press.

Atkins, S., Clear, J. and Ostler, N. 1992. Corpus Design Criteria. *Journal of Literary and Linguistic Computing* 7 (1):1-16.

Atkins, S., Kilgarriff, A., and Rundell, M. 2001. Training workshop in lexicography and lexical computing: course notes. Brighton: University of Brighton.

Baayen, R.H. 2001. *Word Frequencies Distribution*. London: Kluwer Academic Publishers.

Bagwasi, M.M. 2003. The Functional Distribution of Setswana and English in Botswana. *Language, Culture and Curriculum* 16 (2):212-217.

Barnbrook, G. 1996. *Language and Computers; A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Baroni, M. 2006. *Distributions in text.* In Lüdeling, A and Kytö, M. (eds.), Corpus linguistics: An international handbook, Berlin: Mouton de Gruyter.

Baroni, B. and Ueyama, M. 2006. Building general- and special-purpose corpora by Web crawling. Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application. 31-40.

Baugh, S., Harley, A., and Jellis, S. 1996. The role of corpora in compiling the Cambridge International Dictionary of English. *International Journal of Corpus Linguistics* 1 (1):39-59.

Béjoint, H. 2000. *Modern lexicography: An introduction*. New York: Oxford University Press.

Bentivogli, L., Girardi, C. and Piata, E. 2003. The MEANING Italian Corpus. Paper read at the Corpus Linguistics 2003 Conference, Lancaster, UK.

Bentivogli, L. and Pianta, E. 2002. Detecting hidden multiwords in bilingual dictionaries. Paper read at The 10th EURALEX International Conference, at Copenhagen, Denmark.

Bergenholtz, H, and Tarp, S. (eds). 1995. *Manual of specialized lexicography*. Amsterdam, Philadelphia: John Benjamins.

Bharati, A., Prakash, R.K., Sangal R, Bendre, S.M. 2002. Basic statistical analysis of corpus and cross comparison among corpora. Proceedings of International Conference on Natural Language Processing, Mumbai, India. 18-21, December 2002. Available at: http://www.iiit.net/techreports/2002_4.pdf.

Biber, D. 1988. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5 (4):257-269.

Biber, D. 1993. Using register-diversified corpora for general language studies. *Association for Computational Linguistics* 19 (2):219-241.

Biber, D. 1993a. The multi-dimesional approach to linguistic analyses of genre variation; an overview of methodology and findings. *Computers and Humanities* 26:331-345.

Biber, D. 1994. An analytical for framework register studies. In D. Biber and E. Finegan (eds) *Sociolinguistic Perspetives on Register*, Oxford: Oxford University Press.

Biber, D. 1994a. Representative in Corpus Design. *Linguistica Computazionale: Current Issues in Computational Linguistics: In Honour of Don Walker* ix (x):377-407.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison,*. Cambridge: Cambridge University Press.

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

Bindi, R., Calzolari, N., Monachini, M., Pirrelli, V. and Zampolli, A. 1994. Corpora and computational Lexica; integration of different methodologies of lexical knowledge acquisition. *Literary and Linguistic Computing* 9 (1):29-46.

Boldi, P., Codenotti, B., Santini, M., and Vigna, S. 2004. *Structural Properties of the African Web* 2004 [cited 27 April 2004 2004]. Available from http://www2002.org/CDROM/poster/164/.

Borin, L. 2000. A corpus of written Finnish Romani texts. Paper read at Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, 2000, Athens, Greece: ELRA.

Brew, C. and McKelvie, D. 1996. Word-pair extraction for lexicography. Paper read at NeMLaP'96. Ankara, Turkey.

Brown, T.J. 1925. *English Dictionary*. Johannesburg: Pula Press.

Burchfield, R.W. (ed.). 1987. *Studies in lexicography*. Oxford; Oxford University Press.

Burnard, L. 1995. *Users Reference Guide for the British National Corpus version 1.0.* Oxford: Oxford University Press.

Burnard, L. 2002. Where did we go wrong? A retrospective look at the British National Corpus. Paper read at the 4th International conference on teaching and language corpora by doing corpus analysis: Oxford, UK.

Bussman, H. 1996. Routledge Dictionary of language and linguistics. London Routledge.

Butler, C. 1998. Collocational Frameworks in Spanish. *International Journal of Corpus Linguistics* 3 (1):1-32.

Butler, J. 1997. Excitable Speech: A politics of the performative. Routledge: New York.

Campbell, J. 1815. *Travels in South Africa, undertaken at the request of the London Missionary Society: being a narrative of the second journey in the interior of that country*. London: Guthrie.

Cantos, P, and Sanchez, A. 2001. Lexical Constellations: What collocates fail to test. *International Journal of Corpus Linguistics* 6 (2):199-228.

Cavagliá, G. 2005. Measuring the homogeneity and similarity of language corpora. PhD thesis, Information Technology Research Institute, University of Brighton, Brighton.

Čermák, F. 1997. Czech National Corpus: A case in many contexts. *International Journal of Corpus Linguistics* 2 (2):181-197.

Čermák, F. and Křen, M. 2005. Large corpora, lexical frequencies and coverage of text. Paper read at Corpus Linguistics 2005, Birmingham, UK.

Chebanne, A. M. 2002. Glides in Setswana. *Journal of the Linguistics Association for Southern African Development Communities Universities (LASU)* 1:43-49.

Cho, S.W, and O'Grady, W. 1996. Language acquisition: the emergence of a grammar. In W. O'Grady, M. Dobrovolsky and F. Katamba (eds) *Contemporary Linguistics*, London: Longman.

Church, K, and Mercer, R. 1993. Introduction to the special issue on Computational Linguistics using Large Corpora. *Computational Linguistics* 19 (1):1-24.

Church, K. and Hanks, P. 1989. Word Association Norms, Mutual Information, and Lexicography', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*:76-83.

Clear, J. 1987. Computing. In Sinclair, J. (ed). *Looking Up: An Account of the COBUILD Project.*Glasgow: Collins ELT.

Clear, J. 1992. Corpus Sampling. In G. Leiter (ed) *New Directions in English Language Corpora*. New York: Mouton de Gruyter.

Cole, D. 1995. *Setswana-Animals and plants (Setswana-Ditshedi le ditlhare)*. Gaborone: The Botswana Society.

Cole, D.T. 1955. *An Introduction to Tswana Grammar, 9th impression*. Cape Town: Longman.

Cowie, A. P. 1999. English Dictionaries for Foreign Learners: A History. Oxford: Oxford University Press.

Créissels, J., and Chebanne, M.A. 2000. *Dictionnaire Francaise-Setswana, Thanodi Sefora-Setswana*. Mogoditshane: Tasalls Publishing.

Crowdy, S. 1991. *Spoken Corpus Design an Transcription*: Longman.

Crowdy, S. 1993. Spoken Corpus Design. *Literary and Linguistic Computing* 8:259-265.

Crowdy, S. 1994. Spoken Corpus Transcription. *Literary and Linguistic Computing* 9 (1):25-28.

Culpeper, J. 2002. Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. Paper read at Conversations in life and literature ASLA Symposium, at Universitetstryckeriet.

Dash, N.S. and Chaudhuri, B.B. 2000. The process of designing a multidisciplinary monolingual sample corpus, *International Journal of Corpus Linguistics* Vol. 5 (2):179-197.

De Beaugrande, R. 1997. Text linguistics, discourse analysis, and the discourse of dictionaries. *Bibliotheque des cahiers de l'institut de linguistique de Louvain* 87:57-75.

De Haan, P. 1992. The optium corpus sample size. In *Topics in English Linguistics*, edited by G. Leitner. Berlin and New York: Mouton de Gruyter.

De Monnink, I. 1999. Combining corpus with experimental Data. *International Journal of Corpus Linguistics* 4 (1):77-111.

DeRose, S.J. 1991. "An analysis of probabilistic grammatical tagging methods." In *English Computer Corpora: selected papers and research guide*, Stig Johansson and Anna-Brita Stenström (eds). New York: Mouton de Gruyter, pp. 9- 14. Volume 3 of *Topics in English Linguistics*, ed. by Jan Svartvik and Herman Wekker.

De Schryver, G-M. 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic journal of African Studies* 11 (2):266-286.

De Schryver, G-M. 2000 and Prinsloo, D.J. 2000. 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, and Prinsloo, D.J. 2000a. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure. *South African Journal of African Languages* 20 (4):310-330.

De Schryver, G.-M. and Prinsloo D.J. 2000b. *Creating Electronic corpora for the South African Languages*. Available at: http://www.up.ac.za/academic/libarts/afrilang/generalcorpora.htm .

De Schryver, G-M, and Prinsloo, D.J. 2001. Taking Dictionaries for Bantu Languages into the New Millennium – with special reference to Kiswahili, Sepedi and isiZulu in J.S. Mdee & H.J.M. Mwansoko (eds.). 2001. *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam:188–215.

De Schryver, G-M, and Joffe, D. 2004. On how electronic dictionaries are really used. Proceedings of the 11th EURALEX conference on Lexicography, Université de Bretagne Sud, Lorient, France, 6-10 July 2004

Demuth, K, and Johnson, M. 1989. Interactions between discourse functions and agreement in Setawana. *Journal of African languages and linguistics* 11:22-35.

Dent, G. R. 1992. *Compact Setswana Dictionary: English-Setswana, Setswana-English*. 1st ed. Pietermaritzburg: Shutter & Shooter.

Diaz de Ilarraza, A. D, Gurrutxaga, A., Hernaez, A., Lopez de Gerenu, N., and Sarasola, K. 2003. HIZKING21: Integrating language engineering resources and toools into systems with linguistic capabilities. *TALN*:11-14.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1):61-74.

Evert, S. 2004. The Statistics of word coorccurrences: word pairs and collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart.

Evert, S. and Baroni, M. 2005. Testing the extrapolation quality of word frequency models. Paper read at The Corpus Linguistics Conference Series, Burmingham, UK.

Fillmore, C., Ide, N., Jurafsky, D., and Macleod, C. 1998. An American National Corpus: A Proposal. *Proceedings of the First International Language Resources and Evaluation Conference,* Granada, Spain:965-70.

Finch, G. 2000. Linguistic terms and concepts. Basingstoke. Macmillan Press.

Fletcher, W.H. 2002. Making the web more useful as a source for linguistic corpora. Paper read at Corpus linguistics in North America: Selections from Fourth North American Symposium of the American Association for Applied Corpus Linguistics.

Fletcher, W.H. 2005. Concordancing the web: promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus linguistics and the web*. Amsterdam: Rodopi.

Fox, G. 1987. The case for examples. In Sinclair, J. (ed) *Looking up: An account of the COBUILD project in lexical computing*. London: Collins ELT.

Francis, W.N, and Kucera, H. (1964, 1971, 1979). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers*. Providence: Department of Linguistics, Brown University.

Francis, W. N., Kucera, H. and Mackie, A.W. 1982. *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Fredoux, J. 1864. *A sketch of the Sechuana Grammar*. Cape Town.

Fries, U., Tottie, G. and Schneider, P. 1993. The 14th International conference on English language research on computerized corpora. Paper read at Creating and using English Language Corpora, Zurich.

Gaotlhobogwe, M. 2007. *Reteng Refutes Ramsay.* Mmegi Newspaper Online. Available

from http://www.mmegi.bw/2006/May/Wednesday10/75487749879.html.

Garside, R., Leech, G. and Sampson, G. (eds). 1987. *The Computational Analysis of English: A corpus-based approach*. Essex: Longman.

Ghani, R, Jones, R., and Mladenic, D. 2001. Mining the web to create minority language corpora. *CIKM*:279-286.

Ghani, R., Jones, R and Mladenic, D. 2001a. Automatic web search query generation to create minority language corpora. *SIGIR'01*.

Gomez, P.C. 2002. Do we need statistics when we have linguistics? *D.E.L.T.A.* 18 (2):233-271.

Gouws, R.H, and Prinsloo, D.J. 1997. Lemmatisation of Adjectives in Sepedi. *Lexikos* 7:45-57.

Gouws, R.H, and Prinsloo, D.J. 2005. Principles and practices of South African lexicography. *African Sun Media*.

Grefenstette, G. and Tapanainen, P. 1994. What is a word, What is a sentence? Problems of Tokenization. Paper read at Computational Lexicography (COMPLEX '94).

Grimes, B.F. 2004. ETHNOLOGUE: Languages of the World Fourteenth Edition. Available at: *http://www.sil.org/ethnologue/*.

Guenthner, F, and Blanco, X. (in press). Multi-Lexemic Expressions: An Overview. *Lingvisticae Investigationes Supplementa*.

Hanks, P. 1994. Linguistic Norms and Pragmatic Exploitations or, Why Lexicographers need prototype Theory, and Vice Versa. Papers in Computational Lexicography COMPLEX '94, Linguistics Hungarian Academy of Sciences, Budapest.

Hasselbring, S. 2000. *A sociolinguistics survey of the languages of Botswana*. Vol. 1, *Sociolinguistic studies of Botswana language series, Basarwa languages project*. Mogoditshane: Tassals Publishing and Books.

Hasselbring, S., Segatlhe, T and Munch, J. 2001. *A sociolinguistics survey of the languages of Botswana*. Vol. 2, *Sociolinguistic studies of Botswana language series, Basarwa languages project*. Mogoditshane: Tassals Publishing and Books.

Haztivassiloglou, V. 1994. Do we need linguistics when we have statistics? A comparative analysis of the contributions of linguistric cues to statistical word grouping system. In Hundt, M., Nesselhauf, N. and Biewer, C (eds). *The Balancing Art. Combining symbolic and statistical approaches to languages*. Cambridge: The MIT Press.

Heid, U. 1994. Relating lexicon and corpus: computational support for corpus-based lexicon building in DELIS. Paper read at EURALEX, Amsterdam, the Netherlands.

Hinton, P.R. 2004. *Statistics Explained* (2nd ed.). Essex: Routledge.

Hofland, K. and Johansson, S. (1982). Word frequencies in British and American English. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Hornby, A.S. 1996. Oxford Advanced Learner's Dictionary of Current English, 4th impression. Oxford: Oxford University Press.

Horvath, J. 1999. *Advanced Writing in English as a Foreign Language, A corpus-based study of processes and products.* [cited 20 May 2006. Available from http://www.geocities.com/writing_site/thesis/.

Hubert, P., and Labbe, D. 1988. A Model of Vocabulary Partition. *Literary and Linguistic Computing* 3 (4):223-225.

Hull, D. M. 1987. Educational development in Botswana: a plural heritage. *Journal of Negro Education* 56 (3):381-389.

Humphries, C. (ed). 2003. *Philip's Encyclopedia: Comprehensive Edition 2004*. London: Philip's.

Ide, N, and Macleod, C. 2001. The American National Corpus: A Standardized Resource for American English. Paper read at the Corpus Linguistics 2001 conference, at Lancaster University, Lancaster (UK).

Ide, N., Reppen, R, and Suderman, K. 2002. The American National Corpus: more than the web can provide. Paper read at the third Language Resources and Evaluation Conference (LREC), Las Palms, Spain.

Jespersen, O. 1909-49. A modern English grammar on historical principles, I-VII. Copenhagen. Munksgaard.

Johansson, S. 1991. Times change, and so do corpora. In Aijmer, K and Altenburg, B *English corpus linguistics: studies in honour of Jan Svartvikr*. London: Longman.

Johansson, S., and Hofland, K. 1989. *Frequency analysis of English vocabulary and grammar: based on the LOB corpus*. Oxford: Oxford University Press.

Johansson, S., and Stenstrom, A.B. (eds). 1991. *English Computer Corpora: Selected papers and research guide*. In Svartvik, J., and Wekker H. (Eds). *Topics in English Linguistics*. Vol. 3, New York: Mouton de Gruyter.

Johnson, S. 1963. *Johnson's Dictionary: A Modern Selection by E.L. McAdam, Jr. & George Milne*. London: Victor Gollancz.

Jones, D, and Plaatjie, S. 1916/1928. *The Setswana reader the tones of Setswana nouns*. Hants: Gregg International Publishers.

Jones, R, and Ghani, R. 2000. Automatically building a corpus for a minority language from the web. Paper read at 38th Meeting of the ACL, Proceedings of the Student

Research Workshop, Hong Kong.

Keller, F., Lapata, M. and Ourioupina, O. 2002. Using the web to overcome data sparseness. Paper read at the conference on empirical methods in natural language processing (EMNLP), Philadelphia.

Kennedy, G.D. 1998. *An introduction to corpus linguistics*. London; New York: Longman.

Kgasa, M.L.A. 1976. *Thanodi ya Setswana ya Dikole*. Cape Town: Longman.

Kgasa, M.L.A., and Tsonope, J. 1998. *Thanodi ya Setswana*. Gaborone: Longman.

Kilgarriff, A. 1996. Which words are particular of a text? A survey of statistical approaches. Proceedings. AISB Workshop on Language Engineering for Document Analysis and Recognition, 1996, at Sussex University:33-40

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10 (2):135-155.

Kilgarriff, A. 1997a. Using word frequency lists to measure corpus homogeneity and similarity between corpora. Paper read at ACL SIGDAT workshop on very large corpora, Beijing and Hong Kong.

Kilgarriff, A. 2000. Business Models for Dictionaries and NLP. *International Journal of Lexicography* 13 (2):107-118.

Kilgarriff, A. 2001. Comparing Corpora. *International Journal of Corpus Linguistics* 6 (1):97-133.

Kilgarriff, A. 2001. Web as corpus. Paper read at Corpus Linguistics 2001, at Lancaster University (UK).

Kilgarriff, A. 2001b. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. Paper read at Collocation Workshop ACL2001, Toulouse, France.

Kilgarriff, A. 2003. What Computers can and cannot do for lexicography, or Us precision, them recall. Paper read at Asian Association for Lexicography 2003, Urayasu, Chiba, Japan.

Kilgarriff, A., and Grefenstette, G. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3):333-347.

Kilgarriff, A, and Rundell, M. 2001. Lexical Profiling Software and its Lexicographic Applications: A Case Study. Paper read at 10th EURALEX International Congress, Copenhagen, Denmark.

Kilgarriff, A, and Rundell, M. 2002. Lexical profiling software and its lexicographic applications: a case study. Paper read at EURALEX 2002, Copenhagen.

Kilgarriff, A, Rychly, Smrz, P., and D Tugwell. 2004. The Sketch Engine. Paper read at EURALEX 2004, Lorient, France. Available online at: http://www.lexmasterclass.com/people/akcv.htm

Kilgarriff, A, and Salkie, R. 1996. Corpus similarity and homogeneity via word frequency. Paper read at EURALEX '96, at Gothenberg, Sweden.

Kittredge, R. 1982. Variation and Homogeneity of Sublanguages. In Kittredge, R. and Lehrberger, J. (eds) *Sublanguage: Studies of Language in Restricted Semantic Domains*, New York: Walter de Gruyter.

Kjellmer, G. 1994. Lexical differentiators of style: Experiments in lexical variability. Paper read at The Fourteenth International Conference on English Language Research on Computerised Corpora, Zurich.

Kjellmer, G. 1994a. A dictionary of English Collocations, Oxford: Clarendon Press.

Knowles, G., and Don, M.Z. 2004. The notion of a "lemma": Headwords, roots and lexical sets. *International of Journal of Corpus Linguistics* 9 (1):68-81.

Kovarik, J. 2000. How should a large corpus be built? A comparative study of closure in annotated newspaper corpora from two Chinese sources, towards building a larger representative corpus merged from representative sublanguage collections. Paper read at ACL, Hong Kong.

Kucera, H. and Nelson, W. 1965. Computational analysis of present-day American English, Brown University Press

Landau, S.I. 1984. *Dictionaries: the art and craft of lexicography*. New York: Scribner.

Landau, S.I. 1989. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.

Larson, R, and Farber, B. 2006. *Elementary Statistics: Picturing the world*. New Jersey: Pearson Prentice Hall.

Laurén, U. 2002. Some lexical features of immersion pupils' oral and written amd written narration. *Working Papers, Department of Linguistics, Lund University* 50:63-78.

Le Bac, V., Bigi, B., Besacier, L. and Castelli, C.E. 2003. Using the Web for a Fast Language Model Construction in Minority Languages. Paper read at Eurospeech 2003, Geneva.

Lee, D. 2001. Genres, Registers, Text types, Domains, and Styles: Clarifying the Concepts and Navigating a path through the BNC Jungle. *Language Learning and Technology* 5 (3):37-72.

Lee, D. (forthcoming). Computer Corpus-based Linguistics & the uninitiated postgraduate. *to appear in proceedings of the BAAL/CUP seminar: Postgraduate*

research in Applied Linguistics: The Insider Perspective.

Leech, G., Deuchar, M., Hoogenraad, R. 1982. English Grammar for today: a new introduction. Basingstoke. Macmillan Press.

Leech, G. 1991. The state of the art in corpus linguistics. In *English Corpus Linguistics*, edited by K. Aijmer and B. Altenberg. London: Longman.

Leech, G., Rayson, P. and Wilson, A. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Leitner, G. 1992. International Corpus of English: Corpus design - problems and suggested solutions. In *New Directions in English Language Corpora*, edited by G. Leitner. Berling and New York: Mouton de Gruyter.

Leon, J. 2005. Claimed and Unclaimed Sources of Corpus Linguistics. *Henry Sweet Society Bulletin* (44):36-50.

Levin, B., Song, G. and Atkins, S. 1997. Making sense of corpus data: A case study of verbs of sound. *International Journal of Corpus Linguistics* 2 (1):23-64.

Lichtenberk, F. 2003. To list or not to list: Writing a dictionary of a language undergoing rapid and extensive lexical changes. *International Journal of Lexicography* 16 (4):386-401.

Lichtenstein, H. 1928-30. *Travels in Southern Africa in the years 1803, 1804, 1805 and 1806 (a reprint of the translation from the original German by Anne Plumptre)*. Translated by Plumptre, A. Cape Town: The Van Riebeeck Society.

Lichtenstein, H. 1973. *Foundation of the Cape: [and] About the Bechuanas*: A. A. Balkema.

Livingstone, D. 1875-85. *Missionary Travels and Researches in South Africa*. London.

Louw, J.P., and Nida, E.A. 1989. *Greek-English lexicon of the New Testament: based on semantic domains*. New York: United Bible Societies.

Lynch, J (Ed). 2004. *Samuel Johnson's dictionary*. Florida: Levenger Press.

Macleod, C. and Grishman, R. 2000. The influence of corpora on lexicons: corpora use in the creation of COMLEX syntax and NOMLEX. Paper read at 9th EURALEX International Confererence, EURALEX 2000, Universität Stuttgart, Stutthart.

Macleod, C, R Grishman, and A Meyers. 2000 *Dictionaries and balanced corpora: the interdependence of resources* [cited September 2006]. Available from www.nlp.cs.nyu.edu/publication/papers/balanced.ps.

Mair, C. 1992. Problems in the compilation of a corpus of standard Caribbean English: A pilot study. In Leitner, G. (ed). *New Directions in English Language Corpora*. Berlin and New York: Mouton de Gruyter.

Malvern, R, and Richards, B. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19 (1):85-104.

Martin, J.B., and Mauldin, M. 1997. Practical and ethical issues in lexicography: Examples from the Creek dictionary project. In Pye, C. (ed) *1996 Mid-America Linguistics Conference Papers*:565-573.

Mathangwane, J.T. 2002. Reduplicatives and their tonology in Ikalanga. *Journal of the Linguistics Association for Southern African Development Communities Universities (LASU)* 1:50-61.

Matumo, Z.I. 1993. *Setswana English Setswana Dictionary*. Gaborone: Macmillan.

McArthur, T. 1986. *Worlds of reference: lexicography, learning, and language from the clay tablet to the computer*. Cambridge; New York: Cambridge University Press.

McEnery, Tony, Paul Baker, Lou Burnard, and Mark Sebba. 1999. *Mille Working Paper 4: A Comparison of LIDES and TEI Encoding Systems for mark-up of multilingual spoken language data* [cited 20/4/2004]. Available from http://bowland-files.lancs.ac.uk/monkey/ihe/mille/wp4.htm.

McEnery, T, and Wilson, A. 1996. *Corpus linguistics*, *Edinburgh textbooks in empirical linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T, and Xiao, Z. 2003. Fuck revisited. Paper read at Corpus Linguistics 2003, University of Lancaster (UK).

McKee, G., Malvern, D., and Richards, B. 2000. Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Lingusitic Computing* 15 (3):323-337.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Ministry of Education. 1981. *Setswana Standard Orthography of 1981*. Gaborone: Ministry of Education.

Moe, R. 2001. Lexicography and mass production. *Notes on Linguistics* 4 (3):150-156.

Moe, R. 2003. Compiling dictionaries using semantic domains. *Lexikos* 13:215-223.

Moffat, R. 1826. A Bechuana catechism, with translation of the 3rd chapter of the Gospel by John, the Lord's Prayer and other passages of scripture in that language. London.

Moffat, R. 1842. *Missionary Labours and Scenes in Southern Africa*. London: J. Snow.

Molosiwa, A. 2004. Language and literacy issues in Botswana. Michigan: Michigan State University.

Monachini, M, and Picchi, E. 1992. Tagged Corpora: A query system. Paper read at 2nd International conference on computational lexicography, COMPLEX '92, Budapest, Hungary.

Mooko, T. 2004. An investigation into the use of Setswana to teach primary school Mathematics. *Language, Culture and Curriculum* 17 (3):181-195.

Moon, R. 2007. Sinclair, lexicography, and the Cobuild project. *International Journal of Corpus Linguistics* 12:2:159-181.

Murdock, G.P. 1987. *Outline of cultural materials*. 5th ed. New Haven: Relations Area Files.

Naden, T. 1993. From Wordlist to Comparative Lexicography. *Lexikos* 3:167-190.

Nelson, G. 2004. *International Corpus of English: Markup Manual for Written texts* 2002 [cited 04 May, 2004]. Available from http://www.ucl.ac.uk/english-usage/ice/written.pdf.

Nevegina, S.B. 1998. Some problems of borrowing in the Russian language. Вестник Омского университета. 1: 72-75.

Nyati-Ramahobo, L. 1999. *The National language: A resource or a problem - The implementation of the language policy of Botswana*. Gaborone: Pula Press.

Oakes, M.P. 1998. *Statistics for corpus linguistics*, *Edinburgh textbooks in empirical linguistics*. Edinburgh: Edinburgh University Press.

Onibere, E. Morgan, A.S., Busang, E.M. and Mpoeleng, D. 2001. Human-computer interfaces design isues for a multi-lingual English speaking country - Botswana. *Interacting with Computers* 13:497-512.

Ooi, V.B.Y. 1998. *Computer corpus lexicography*, *Edinburgh textbooks in empirical linguistics.* Edinburgh: Edinburgh University Press.

Oostdijk, N. 1988. A corpus linguistics approach to linguistic variation. *Literary and linguistic Computing* 3 (1):12-25.

Otlogetswe, T. 2004. The BNC Design as a Model for a Setswana Language Corpus. *Proceedings of CLUK'04, Birmingham, UK*:193-198.

Otlogetswe, T. 2006. Challenges to issues of balanace and representativeness in African lexicography. *Lexikos* 16:145-160.

Pagano, R. 2001. *Understanding Statistics in the behavioral sciences* (6th ed.). Belmont: Wadworth, Thomson Learning.

Paikeday, T.M. 1992. O Corpora! *Lexicographica*:307-317.

Pearsall, J. 1998. *The New Oxford Dictionary of English*. Oxford: Oxford University

Press.

Peters, M.A. 1982. *Bibliography of the Tswana language*. Pretoria: State Library.

Poplack, S. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French Project. In Fasold, R.W. and Schriffrin, D. (eds). *Language Change and Variation*, Amsterdam, Philadelphia: John Benjamins Publishing Company.

Pravec, N.A. 2002. Survey of learner corpora. *ICAME Journal* 26:81-114.

Prinsloo, D.J. 2004. Revising Matumo's Setswana-English-Setswana Dictionary. *Lexikos* 14:158-172.

Prinsloo, D.J. and De Schryver, G-M. 1999. The lemmatization of nouns in African languages with special reference to Sepedi and Ciluba. *South African Journal African Languages* 19 (4):258-275.

Prinsloo, D.J., and De Schryver, G-M. 2001. Monitoring the Stability of a Growing Organic Corpus, with special reference to Sepedi and Xitsonga. *Dictionaries: Journal of The Dictionary Society of North America 22*:85-129.

Prinsloo, D.J., and Gouws, R.H. 1996. Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho. *South African Journal of African Languages* 16 (3):100-107.

Ramsay, J. 2006. *Botswana's ethnic breakdown – OP*. Online Mmegi Newspaper [cited 22 August 2006. Available at http://www.mmegi.bw/2006/May/Tuesday9/754876171530.html.

Rayson, P. and Garside, R. 2000. Comparing corpora using frequency profiling. In proceedings of the *workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. 1-8 October 2000, Hong Kong:1-6.

Rayson, P, Berridge, D. and Francis, B. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. Paper read at 7th International Conference on Statistical analysis of textual data, at Louvain-la-Neuve, Belgium:926-936.

Rayson, P, Leech, G. and Hodges, M. 1997. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2 (1):133-152.

Rayson, P., Wilson, A. and Leech, G. 2002. Grammatical word class variation with the British National Corpus Sampler. Paper read at New Frontiers of Corpus Research: 21st International Conference on English Language Research on Computerized Corpora, Sydney.

Rayson, P. 2003. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis. Computer Science, Computing Department,

Lancaster University, Lancaster.

Reményi, A.A. 2001. Use logbooks and find the original meaning of "representativeness". Paper read at The Corpus Linguistics 2001, at UCREL, Lancaster.

Renouf, A. 1987. Corpus Development. In J. Sinclair (ed) *Looking Up: An account of the COBUILD*, London, Glasgow: COBUILD.

Republic of Botswana. 1994. *Government Paper No. 2*: *Revised National Policy on Education* Gaborone, Botswana: Government Printer.

Republic of Botswana. 2001. Central Statistics Office. 2001 census. http://www.cso.gov.bw/

Roberts, G.F. 1998. The home page as genre: a narrative approach. Paper read at Annual Hawaii International Conference on System Sciences, at Hawaii.

Roget, P.M. 1958. *Roget's Thesaurus*. Harmondsworth: Middlesex: Penguin Books.

Rundell, M. 1996. *The corpus of the future and the future of the corpus* Talk at niversity of Exeter, special conference on 'New Trends in Reference Science', [cited 17 July 2004. Available from http://www.ruf.rice.edu/~barlow/futcrp.html.

Salt, H. 1814. *A voyage to Abyssinia and travels into the interior of that country*. London.

Sanchez, A., and Cantos, P. 1997. Predictability of word forms (types) and lemmas in linguistic corpora. A case study on the analysis of the CUMBRE corpus: An 8-million-word corpus of contemporary Spanish. *International Journal of Corpus Linguistics* 2 (2):259-280.

Sandilands, A. 1953. *Introduction to Tswana*. Tigerkloof: The London Missionary Society.

Santini, M. 2003. Identifying genres on the web. Departmental Report, ITRI-03-06, Brighton: University of Brighton.

Sardinha, B. 1999. Using key words in text analysis: practical aspects. DIRECT Papers 42:1-9

Sardinha, B. 1999a. Beginning Portuguese corpus linguistics: exploring a corpus to teach Portuguese as a foreign language. *D.E.L.T.A.* 15 (2):289-299.

Sardinha, T.B. 2000. Comparing corpora with wordsmith keywords: Comparação de corpora com WordSmith KeyWords. the ESP, São Paulo, vol. 22, no 1:187-99

Savage, D. 1990. *Testing language attitudes and use*. In Bergman, T.G. (ed), *Survey reference manual*. Dallas: Summer Institute of Linguistics.

Scannell, K.P. 2007. "The Crubadan Project: Corpus building for under-resourced languages". *Cahiers du Central*, 5: 1-10.

Scott, M. 2004-2006. Oxford WordSmith Tools version 4. Oxford: Oxford University Press.

Selolwane, O.D. 2004. Ethnic structure, inequality and governance of the public sector in Botswana. *UNRISD Project on Ethnic Structure, Inequality and Governance of the Public Sector*.

Setati, M., Adler, J., Reed, Y. and Bapoo, A. 2002. Incomplete journeys: Code-switching and other language practices in Mathematics, Science and English language classrooms in South Africa. *Language and Education* 16 (2):128-149.

Sharroff, S. 2004. Methods and Tools for development of the Russian Reference Corpus. In Archer, D., Wilson, A., and Rayson, P. *Corpus Linguistics around the World*, Amsterdam: Rodopi.

Sharoff, S. 2006. Creating general purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S. (eds). WaCky! Working papers on the Web as Corpus. Gedit, Bologna. Also: http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf

Shepherd, M, and Watters, C. 1998. The evolution of cybergenres. Paper read at 31st Annual Hawaii international Conference on System Sciences, at Hawaii.

Sigley, R. 1997. Text Categories and Where You can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics* 2 (2):199-237.

Simons, G. F. 1998. In search of task-centered software: building single purpose tools from multipurpose components: SIL Electronic working papers 1998-004.

Sinclair, J. 1989. Corpus Creation. In *Language, Learning and Community*, edited by Candlin and McNamara: NCELTR Macquire University.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1996. EAGLES. Preliminary Recommendations on Corpus Typology. Available at http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html. (Accessed 16.09.2006).

Sinclair, J. 2005. Corpus and Text - Basic Principles [cited 20 October 2005]. In *Developing Linguistic Corpora: a guide to good practice*, edited by M. Wynne. Oxford: Oxbow Books, available online from http://ahds.ac.uk/linguistic-corpora/.

Snyman, J.W., J.S. Shole, and J.C. Le Roux. 1990. *Dikisinare ya Setswana English Afrikaans Dictionary. Woordeboek.* Pretoria: Via Afrika Limited.

Southerland, R. H, and Katamba F. 1996. Language in social contexts. In W. O'Grady,

M. Dobrovolsky and F. Katamba (eds). *Contemporary Linguistics: An Introduction*. London, New York: Longman.

Sperberg-McQueen, C.M. and Burnard, L. (eds). 1994. *Guidelins for electronic text encoding and interchange*. Chicago and Oxford: Text Encoding Initiative.

Svensén, B. 1993. *Practical Lexicography*. Oxford: Oxford University Press.

Summers, D. 1995. *Longman Dictionary of Contemporary English.* Longman, Essex.

Summers, D. 1996. Computer lexicography: the importance of representativeness in relation to frequency. In Thomas, J. and Short, M. (eds). *Using corpora for language research*. London: Longman:260-266.

Summers, D. (ed). 1997. Longman Essential Activator. Harlow, Essex: Addison Wesley Longman.

Teubert, W. 2001. Corpus linguistics and lexicography. *International Journal of Linguistics* 6 (special issue):125-153.

Thekiso, E. 2001. A Sociolinguistic Analysis of Communication Processes in a Bilingual Court of Law in Gaborone, Botswana, PhD Thesis, University of Warwick, UK.

Thomson, N. 1989. How to read Articles which Depend on Statistics. *Literary and Linguistic Computing* 4 (1):6-11.

Thorndike, E. 1921. The Teacher's Word Book. New York: Teachers College.

Van Warmelo, N.J. 1931. *Kinship Terminology of the South African Bantu. Pretoria*. Pretoria: Government Printer.

Varadi, T. 2001. The linguistic relevance of corpus linguistics. Paper read at Corpus Linguistics 2001, at Lancaster University.

Verlinde, S. and Selva, T. 2001. Corpus-based versus intuition-based lexicography: defining a wordlist for a French learner's dictionary. Paper read at Corpus Linguistics 2001, at UCREL, Lancaster University UK.

Villasenor-Pineda, L., M. Montes-y-Gómez, M. Pérez-Coutino, and D. Vaufreydaz. 2003. A corpus balancing method for language model construction. *Fourth International Conference on Intelligent Text Processing and Computational Linguistics*:393-401.

Vintar, Š. 1999. A Lexical Analysis of the IJS-ELAN Slovene-English Parallel Corpus. In: Vintar, Š. (ed.) *Proceedings of the workshop Language Technologies – Multilingual Aspects*. Ljubljana: Faculty of Arts, 63-69.

Volk, M. 2002. Using the web as corpus for linguistic research. In *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, edited by R.

Pajusalu and T. Hennoste. University of Tartu: Department of General Linguistics.

Wells, R. A. 1973. *Dictionaries and the Authoritarian Tradition*. The Hague: Mouton & Co. N.V., Publishers.

Wierzbicka, A. 1985. *Lexicography and conceptual analysis*. Ann Arbor: Karoma Publishers.

Wilson, A. and Rayson, P. (1993). Automatic Content Analysis of Spoken Discourse. In: C. Souter and E. Atwell (eds), Corpus Based Computational Linguistics. Amsterdam: Rodopi:215-226.

Wisker, G. 2001. *The postgraduate research handbook*. New York: Palgrave.

Wookey, A.J. 1904. *Setswana and English phrases with short introduction to grammar and a vocabulary*. Cape Town: Townshend & Son.

Wynne, M. (ed). 2005. *Developing Linguistic Corpora; a guide to good practice*. Oxford: Oxbow books, available online from http://ahds.ac.uk/linguistic-corpora/.

Xiao, Z., and A. McEnery. 2005. Two approaches to genre analysis. *Journal of English Linguistics* 33 (1):62-82.

Youmans, G. 1990. Measuring lexical style and competence: The type-token vocabulary curve. *Style* 24:584-599.

Zgusta, L. 1971. *Manual of lexicography*. [Scientific ed, *Janua linguarum. Series maior, 39*. The Hague, Mouton.

Zipf, G. K. 1949. *Human behavious and the principle of least effort*. Cambridge MA: Addison-Wesley.

Zipf, G. K. 1965. *The psycho-biology of language*. Cambridge MA: MIT Press.

Zwicky, A.M., and A.D. Zwicky. 1982. Registers as a Dimension of Linguistic Variation. In in Kittredge, R. and Lehrberger, J. (eds). *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin, New York: Walter de Gruyter.

Appendix 1: Proposed subentries of *pelo* headword

1. ama pelo,
2. balabala ka pelo,
3. baya pelo,
4. beta pelo,
5. betwa ke pelo,
6. bofa pelo,
7. bolawa ke pelo,
8. bolwetse jwa pelo,
9. bona pelo,
10. bongwefela jwa pelo,
11. bonosi jwa pelo,
12. boteng jwa pelo,
13. bua ka pelo,
14. bula pelo,
15. busa pelo,
16. fela pelo,
17. feretlha pelo,
18. fetola pelo,
19. gapa pelo,
20. garoga pelo,
21. kgaoga pelo,
22. go sena letsapa le fisang pelo,
23. gonolwa ke pelo,
24. isa pelo mafisa,
25. itaya pelo,
26. itse pelo,
27. kgwaralatsa pelo,
28. lala ka pelo e rotha madi,
29. mabetwa-e-pelo,
30. masetla pelo,
31. matlhomola pelo,
32. matlhotlha-pelo,
33. nametsa pelo,
34. ngomola pelo,
35. ngona pelo,
36. nna pelo,
37. nona pelo ka mathe,
38. ntsha pelo,
39. ntsha pelo pelaelo,
40. pateletsa pelo,
41. pelo e boela mannong,
42. pelo e e botlhoko,
43. pelo e e letlapa,
44. pelo e ja serati,
45. pelo e khibidu,
46. pelo e rotha madi,
47. pelo e rutha,
48. pelo e setlhogo,
49. pelo e thata,
50. pelo khutshwane,
51. pelo namagadi,
52. pelo ntsho,
53. pelo pedi,
54. pelo pholwana e a golegwa
55. pelo potsane e a golegwa
56. pelo tshweu,
57. pelo yotlhe,
58. pelo-e-thata,
59. pelo-kgale,
60. pelo-telele,
61. pelo-tlhomogi,
62. pelo-tshetlha,
63. phatlola pelo,
64. ritibatsa pelo,
65. sephiri sa pelo,
66. sera pelo,
67. sethunya sa pelo,
68. sisa pelo,
69. sulafatsa pelo,
70. swa pelo,
71. swegaswega pelo,
72. thiba maroba a pelo,
73. thuba pelo,
74. tlala pelo,
75. tlalelana pelo,
76. tlhomola pelo,
77. tlola pelo,
78. tshwara ka pelo,
79. tshwara pelo,
80. tswa pelo,
81. tswela pelo,
82. uba pelo,
83. wa pelo,
84. wela pelo

Appendix 2: Participation consent form

UNIVERSITY OF BOTSWANA & UNIVERSITY OF BRIGHTON

Recording Speech

Thank you very much for agreeing to take part in this project. The study is being carried out by Thapelo Otlogetswe, a lecturer in the Department of English of the University of Botswana currently pursuing doctoral studies with the University of Brighton, UK. His research will go a long way in compiling a national treasury of the Setswana language which will inform Setswana dictionary writers and language researchers on how words are used in ordinary, everyday conversation. This resource will provide a record of how the Setswana language is currently spoken.

We are asking a large cross-section of people around the country to help with this task by recording their own conversations. These will then be transcribed on computer and built into a database which will contain several million words, and will be used for language research.

Confidential information like personal names, phone numbers and addresses will be deleted from the tapes and transcripts.

What we would like you to do is to record your conversations using the personal stereo provided. You will also need to write down some details of all conversations you have in the forms provided.

If you have any problems with recording or filling in the form ring Thapelo Otlogetswe on 71859452 or the Secretary of the Department of English at 355 2624 who will be able to help you.

- I agree to take part in this research which is to record my conversation with others.
- I am aware that my recorded conversation will become part of a collection of texts that will be used research.
- I understand that I am free to withdraw from the investigation at any time.

Name (please print) ……

Signed ……

Date …….

OTHER SPEAKERS ON THE TAPE SHOULD GRANT APPROVAL TO BE RECORDED BY WRITING THEIR NAMES AND SIGNING BELOW.

1. ……
2. ……
3. ……
4. ……
5. ……
6. ……

Appendix 3: Conversation log

On this page please write in details of conversations recorded on TAPE NO____ SIDE____

Date started recording on this side of the tape ___/ ___ /___ (e.g. 30th January 2003 = 30/01/03)

Time started recording on this side of the tape ___ am/pm

Conversation took place in _____ cattlepost/lands/village/city

What were you doing during the conversations? – e.g. paperwork at work, cooking at home, relaxing at home, travelling by bus etc. THIS INFORMATION IS FOR USE BY THE RESEARCHER ONLY.

WRITE BELOW WHAT YOU WERE DOING WHILE RECORDING ON THIS SIDE OF TAPE.

In the space below please write in the first names or initials and details (where you know them) of the people speaking on this side of the tape. Do not forget to include YOUR OWN details.

	FIRST NAME	OCCUPATION	AGE	SEX	TRIBE	RELATIONSHIP to yourself (wife, son, friend)
1						
2						
3						
4						

Mo tsebeng e, kwala dintlha ka kgatiso e o e dirileng mo khaseteng ya nomore ____

lotlhakore____
Kgwedi le letsatsi tse kgatiso e similotseng ka tsone __/ ___/ ___ (sekai: 30 Ferikgong 2003 = 30/01/03)
Nako e kgatiso e simolotsweng ka yone ____ am/pm

Puisanyo e e gatisitswe kwa morakeng/masimong/motse/toropo ya _____

O ne o dira eng fa o gatisa puisanyo e? (sekai: o theogetse mo ofising, o iketlile kwa gae, o le mo baseng, o tlhatswa, jalo jalo)

KWALA FA TLASE SE O NENG O SE DIRA FA O GATISA LETLHAKORE LE LA KHASETE.

Fa tlase kwala maina le tse dingwe ka ga batho ba ba buiwang mo lotlhakoreng lo lwa khasete. O seka wa lebala go kwala leina la gago.

	LEINA LA NTLHA	TIRO	DINGWAGA	BONG	MORAFE	KAMANO ya gago le babui (monnao, tsala)
1						
2						
3						
4						

Appendix 4: Headteacher's letter

The Headteacher
XXXX Primary School
P. O. Box XXX
Mochudi
Botswana

19 August 2004

Dear Sir/Madam

RE: REQUEST FOR RECORDING SPOKEN SETSWANA

I am writing to request permission to record classroom interaction in Setswana classes in your school. These interactions will be transcribed for inclusion in the creation of a Setswana Corpus which will be analysed as part of doctoral studies. A corpus is a collection of texts in a computer (or digital form) for linguistic analysis.

I am a Linguistics lecturer in the Department of English at the University of Botswana. Currently I am pursuing doctoral studies at the Information Technology Research Institute, University of Brighton, UK. My research is in CORPUS LEXICOGRAPHY – the use of huge language databases for the study and creation of dictionaries and dictionary resources.

The nature of my research requires varieties of texts running into millions of words. These texts are usually obtained from newspapers, magazines, conversations, novels, plays, speeches, radio news, classroom interactions and many other sources. To build such a database, I have so far received texts from Macmillan [about 1 million words], *Mokgosi* newspaper [over 1 million words], *Naledi* newspaper [*Mmegi* newspaper insert], Department of Information and Broadcasting, and from different departments in Botswana. I am therefore making an appeal to you, that I come and record Setswana classroom interactions in your school.

The recording of Setswana classroom interactions is part of a larger project of capturing spoken Setswana in which a large cross-section of people around the

country is taking part by recording their own conversations. These will then be transcribed on computer and built into a database which will contain several million words, and will be used for scientific research which will inform PhD research and in the long term improve dictionaries and language research.

Such a study raises some ethical issues. To take care of these issues, recordings from schools will be stored in a computer completely anonymously. No one will know who has used which words. Names of teachers, students or schools will not be entered into the computer and do not have to be recorded. There will be no association between any school and any recordings to protect the integrity and privacy of the school, especially that such an association is irrelevant to the study. What will be recorded, however, is the region from which the recording is done [e.g. Southern]; the type of school [primary or secondary], the level or class of students [e.g. std 6 or form 5] and the regional origins of the teacher [e.g. Southern, Kweneng, Central, or South-East]. All participating schools in this research will be suitably acknowledged for aiding research in the Setswana language.

These recordings and transcriptions together will provide a record of how the Setswana language is spoken currently.

Any enquiry or query related to this research should be left with the Secretary to the Department of English, University of Botswana at: Tel: 355 2624. Alternatively, please fax any question or query to: 5985 098, or email me at thapelo.otlogetswe@itri.brighton.ac.uk or otlogets@mopipi.ub.bw.
You are also welcome to contact me directly at the following address:

THAPELO J. OTLOGETSWE
University of Botswana
Department of English
Private Bag UB 00703
Gaborone

I hope you will be able to grant us this permission and in so doing aid the development of an important national resource. Please find attached information outlining the logistics of executing the recording process.

Yours sincerely

_____

Appendix 5: Accompanying details for classroom recordings

This page briefly outlines the process to be taken in making the proposed recordings of Setswana language classes and how such recordings may be attempted without disrupting the smooth running of teaching and learning.

We propose and envisage the following:

1. That details of the proposed research should be discussed between Setswana teacher(s) and the School Head before a trip for recording is made to the school.
2. That the researcher establishes telephonic contact with the School Head or the delegated person to agree on a specific day to travel to the school.
3. We propose to make recordings on a day between mid-September to mid-November.
4. The researcher comes to the school and establishes contact with the School Head and the relevant teachers, or at least some of them.
5. That as much as possible all the recording in a school should be completed in a single day.
6. Teachers should record their teaching themselves.
7. The researcher does not have to go into any classroom. We hope this will reduce tension on the part of the teacher since they will not feel observed.
8. A personal recorder will be provided to the teacher to take to class to record themselves.
9. At least 5 instances of teaching in a school should be recorded.
10. It is important to emphasise that this research will in no way make judgements of a pedagogical nature. We will not make judgements of whether the content delivered is relevant or judgements on a teacher's voice projection. Such considerations fall outside the remit of this research. What is important is that the Setswana language is used. What is taught and how it is taught, is not significant to this research. It must be emphasized that what will be quantified in this study is words and their usage only.

Any enquiry or query related to this research may be left with the Secretary to the Department of English, University of Botswana at: Tel: 355 2624. Alternatively

please fax any question or query to: 5985 098, or email me at thapelo.otlogetswe@itri.brighton.ac.uk or otlogets@mopipi.ub.bw. You are also welcome to contact me directly at the following address:

THAPELO J. OTLOGETSWE
University of Botswana
Department of English
Private Bag UB 00703
Gaborone

I will be happy to discuss details of this research with you or any teacher who may want to find out more.

With many thanks

_____

THAPELO J. OTLOGETSWE [MR.]

Appendix 6: Letter to publishers asking for text

Dear Sir/Madam

RE: REQUEST FOR SETSWANA WRITTEN TEXT

I am writing to request permission to access your Setswana written classes in your school. The Setswana text will be included in the creation of a Setswana corpus which will be analysed as part of doctoral studies. A corpus is a collection of texts in a computer (or digital form) for linguistic analysis. The corpus will also aid Setswana research beyond PhD research.

I am a Linguistics lecturer in the Department of English at the University of Botswana. Currently I am pursuing doctoral studies at the Information Technology Research Institute, University of Brighton, UK. My research is in CORPUS LEXICOGRAPHY – the use of huge language databases for the study and creation of dictionaries and dictionary resources.

The nature of my research requires varieties of texts running into millions of words. These texts are usually obtained from newspapers, magazines, conversations, novels, plays, speeches, radio news, classroom interactions and many other sources. To build such a database, I have so far received texts from Macmillan [about 1 million words], *Mokgosi* newspaper [over 1 million words], *Naledi* [*Mmegi* newspaper insert], Department of Information and Broadcasting, and from different departments in Botswana. I am therefore making an appeal to you for more Setswana text.

The Setswana text that we are requesting will be part of a larger project of capturing the use of varieties of Setswana

Any enquiry or query related to this research should be left with the Secretary to the Department of English, University of Botswana at: Tel: 355 2624. Alternatively, please fax any question or query to: 5985 098, or email me at thapelo.otlogetswe@itri.brighton.ac.uk or otlogets@mopipi.ub.bw.
You are also welcome to contact me directly at the following address:

THAPELO J. OTLOGETSWE

University of Botswana
Department of English
Private Bag UB 00703
Gaborone

I hope you will be able to grant us access to the text in so doing aid the development of an important national resource. Please find attached information outlining the logistics of executing the recording process.

Yours sincerely

_____

THAPELO J. OTLOGETSWE [MR.]

Appendix 7: BNC Part-of-speech codes

(from www.kilgariff.co.uk)

AJ0

adjective (general or positive) e.g. *good, old*

AJC

comparative adjective e.g. *better, older*

AJS

superlative adjective, e.g. *best, oldest*

AT0

article, e.g. *the, a, an, no* . Note the inclusion of *no*: articles are defined as determiners which typically begin a noun phrase but cannot appear as its head.

AV0

adverb (general, not sub-classified as **AVP** or **AVQ**), e.g. *often, well, longer, furthest*. Note that adverbs, unlike adjectives, are not tagged as positive, comparative, or superlative. This is because of the relative rarity of comparative or superlative forms.

AVP

adverb particle, e.g. *up, off, out*. This tag is used for all prepositional adverbs, whether or not they are used idiomatically in phrasal verbs such as "Come out here", or "I can't hold out any longer".

AVQ

wh-adverb, e.g. *when, how, why*. The same tag is used whether the word is used interrogatively or to introduce a relative clause.

CJC

coordinating conjunction, e.g. *and*, or, *but*.

CJS

subordinating conjunction, e.g. *although, when*.

CJT

the subordinating conjunction *that*, when introducing a relative clause, as in "the day that follows Christmas". Some theories treat *that* here as a relative pronoun; others as a conjunction. We have adopted the latter analysis.

CRD

cardinal numeral, e.g. *one, 3, fifty-five, 6609*.

DPS

possessive determiner form, e.g. *your, their, his*.

DT0

general determiner: a determiner which is not a **DTQ** e.g. *this* both in "This is my house" and "This house is mine". A *determiner* is defined as a word which typically occurs either as the first word in a noun phrase, or as the head of a noun phrase.

DTQ

wh-determiner, e.g. *which, what, whose, which*. The same tag is used whether the word is used interrogatively or to introduce a relative clause.

EX0

existential *there*, the word thereappearing in the constructions "there is...", "there are ...".

ITJ

interjection or other isolate, e.g. *oh, yes, mhm, wow*.

NN0

common noun, neutral for number, e.g. *aircraft, data, committee*. Singular collective nouns such as *committee* take this tag on the grounds that they can be followed by either a singular or a plural

verb.

NN1

singular common noun, e.g. *pencil, goose, time, revelation*.

NN2

plural common noun, e.g. *pencils, geese, times, revelations*.

NP0

proper noun, e.g. *London, Michael, Mars, IBM*. Note that no distinction is made for number in the case of proper nouns, since plural proper names are a comparative rarity.

ORD

ordinal numeral, e.g. *first, sixth, 77th, next, last*. No distinction is made between ordinals used in nominal and adverbial roles. *next* and *last* are included in this category, as general ordinals.

PNI

indefinite pronoun, e.g. *none, everything, one* (pronoun), *nobody*. This tag is

applied to words which always function as heads of noun phrases. Words like *some* and *these*, which can also occur before a noun head in an article-like function, are tagged as determiners, **DT0** or **AT0**.

PNP

personal pronoun, e.g. *I, you, them, ours*. Note that possessive pronouns such as *ours* and *theirs* are included in this category.

PNQ

wh-pronoun, e.g. *who, whoever, whom*. The same tag is used whether the word is used interrogatively or to introduce a relative clause.

PNX

reflexive pronoun, e.g. *myself, yourself, itself, ourselves*.

POS

the possessive or genitive marker *'s* or *'*. Note that this marker is tagged as a distinct word. For example, "Peter's or someone else's" is tagged

PRF

the preposition *of*. This word has a special tag of its own, because of its high frequency and its almost exclusively postnominal function.

PRP

preposition, other than *of*, e.g. *about, at, in, on behalf of, with*. Note that prepositional phrases like *on behalf of* or *in spite of* are treated as single words.

TO0

the infinitive marker *to*.

UNC

"unclassified" items which are not appropriately classified as items of the English lexicon. Examples include foreign (non-English) words; special typographical symbols; formulae; hesitation fillers such as *errm* in spoken language.

VBB

the present tense forms of the verb *be*, except for *is* or *'s am, are 'm, 're, be* (subjunctive or imperative), *ai (as in ain't*).

VBD

the past tense forms of the verb *be, was, were*.

VBG

-ing form of the verb *be, being*.

VBI

the infinitive form of the verb *be, be*.

VBN

the past participle form of the verb *be, been*

VBZ

the -s form of the verb *be, is, 's*.

VDB

the finite base form of the verb *do*, *do*.

VDD

the past tense form of the verb *do*, *did*.

VDG

the -ing form of the verb *do*, *doing*.

VDI

the infinitive form of the verb *do*, *do*.

VDN

the past participle form of the verb *do*, *done*.

VDZ

the -s form of the verb *do*, *does*.

VHB

the finite base form of the verb *have*, *have, 've*.

VHD

the past tense form of the verb *have, had, 'd*.

VHG

the -ing form of the verb *have, having*.

VHI

the infinitive form of the verb *have*, *have*.

VHN

the past participle form of the verb *have*, *had*.

VHZ

the -s form of the verb *have*, *has, 's*.

VM0

modal auxiliary verb, e.g. *can, could, will, 'll, 'd, wo*(as in *won't*)

VVB

the finite base form of lexical verbs, e.g. *forget, send, live, return*. This tag is used for imperatives and the present subjunctive forms, but not for the infinitive (VVI).

VVD

the past tense form of lexical verbs, e.g. *forgot, sent, lived, returned*.

VVG

the -ing form of lexical verbs, e.g. *forgetting, sending, living, returning*.

VVI

the infinitive form of lexical verbs , e.g. *forget, send, live, return*.

VVN

the past participle form of lexical verbs, e.g. *forgotten, sent, lived, returned*.

VVZ

the -s form of lexical verbs, e.g. *forgets, sends, lives, returns*.

XX0

the negative particle *not* or *n't*.

ZZ0

alphabetical symbols, e.g. *A, a, B, b, c, d*.

The following *portmanteau tags* are used to indicate where the CLAWS system has indicated an uncertainty between two possible analyses:

AJ0-AV0

adjective or adverb

AJ0-NN1

adjective or singular common noun

AJ0-VVD

adjective or past tense verb

AJ0-VVG

adjective or -ing form of the verb

AJ0-VVN

adjective or past participle

AVP-PRP

adverb particle or preposition

AVQ-CJS

wh-adverb or subordinating conjunction

CJS-PRP

subordinating conjunction or preposition

CJT-DT0

that as conjunction or determiner

CRD-PNI

one as number or pronoun

NN1-NP0

singular common noun or proper noun

NN1-VVB

singular common noun or base verb form

NN1-VVG

singular common noun or -ing form of the verb

NN2-VVZ

plural noun or -s form of lexical verb

VVD-VVN

past tense verb or past participle

The following codes are used with c elements only:

PUL

left bracket (i.e. (or [ )

PUN

any mark of separation (. ! , : ; - ? ...)

PUQ

quotation mark (` ' `` ")

PUR

right bracket (i.e.) or])

Note that some punctuation marks (notably long dashes and ellipses) are not tagged as such in the corpus, but appear simply as entity references.