Bibliography


Research Workshop, Hong Kong.


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.


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.


research in Applied Linguistics: The Insider Perspective.


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.


Lancaster University, Lancaster.


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


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


Appendix 1: Proposed subentries of *pelo* headword

1. ama pelo, 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 seithogo, 49. pelo e thata, 50. pelo khotshwane, 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-thomogi, 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. thomola 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.

<table>
<thead>
<tr>
<th>FIRST NAME</th>
<th>OCCUPATION</th>
<th>AGE</th>
<th>SEX</th>
<th>TRIBE</th>
<th>RELATIONSHIP to yourself (wife, son, friend)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mo tsebeng e, kwala dintlha ka kgatiso e o e dirileng mo khaseteng ya nomore _____
lotlhakore____
Kgwedi le letsatsi tse kgatiso e similotseng ka tsoné __/___/___ (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.

<table>
<thead>
<tr>
<th>LEINA NTLHA</th>
<th>LA TIRO</th>
<th>DINGWA GA</th>
<th>BONG</th>
<th>MORAFE</th>
<th>KAMANO ya gago le babui (monnao, tsala)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
THAPELO J. OTLOGETSWE [Mr.]
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 the email addresses provided. 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.
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 there appearing 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.