

Corpus-based Lexicographic Pragmatics: On 'transforming' dirty corpora

Gilles-Maurice de Schryver

Ghent University, Belgium;
University of the Western Cape, South Africa; and
TshwaneDJe HLT

Corpus-based pragmatics in dictionaries

Just over one decade ago, corpus-based pragmatics labelling made its way into a major dictionary for the first time. In the *Collins COBUILD English Dictionary* (COBUILD 2, Sinclair 1995), the compilers started to insert the novel 'PRAGMATICS' sign into the Extra Column, whenever the 'statement of meaning' for certain senses of words had to be supplemented by an 'added meaning'. In the latest edition, COBUILD 5, this single label is split into seven different labels, as can be seen from Figures 1 and 2.

Pragmatics

People use language to achieve different goals – they invite, give compliments, give warnings, show their emotions, tell lies, and make commitments. The ability to use language effectively to fulfil intentions and goals is known as pragmatic competence, and the study of this ability is called pragmatics. The analysis of language which has been used to prepare this dictionary is based on the idea that speakers and writers plan and fulfil goals as they use language. This in turn entails choices. Speakers choose their goals and they choose appropriate language for their goals.

Different languages use different pragmatic strategies. In order to use a language effectively, and be successful in achieving your goals, you need to know what the pragmatic conventions are for that particular language. It is therefore important that learners of English are given as much information as possible about the ways in which English speakers use their language to communicate.

Because of the large amounts of data in The Bank of English®, COBUILD is uniquely placed to help learners with pragmatics. We have analyzed the data and have found, for example, the ways in which English speakers express approval and disapproval, show their emotions, or emphasize what they are saying.

In the dictionary, we draw attention to certain pragmatic aspects of words and phrases of English, paying special attention to those that, for cultural and linguistic reasons, we feel may be confusing to learners. We do this by having a label in the Extra Column to show the type of pragmatic information being given. The following labels are used in the dictionary.

approval
You can choose words and expressions to show that you approve of the person or thing you are talking about, e.g. *angelic*.

disapproval
You can choose words and expressions to show that you disapprove of the person or thing you are talking about, e.g. *brat*.

emphasis
Many words and expressions are used to emphasize the point you are making, e.g. *never-ending*.

feelings
Another function of pragmatics is to express your feelings about something, or towards someone, e.g. *unfortunately*.

formulae
There are many words and expressions in English which are fairly set, and are used in particular situations such as greeting and thanking people, or acknowledging something, e.g. *hi*, *congratulations*.

politeness
Certain words and expressions in English are used to express politeness, sometimes even to the point of being euphemistic, e.g. *elderly*.

vagueness
Speakers and writers use many words and expressions in English to show how certain they are about the truth or validity of their statements. We have called this type of pragmatic information 'vagueness', though it is sometimes also called 'hedging' or 'modality', e.g. *presumably*.

We hope that you will enjoy learning about pragmatics in the English language. Pragmatics, in any language, is central to communication. When you can understand the context and subtle meanings of a word, you can give and receive accurate messages. This should enable you to achieve your pragmatic goals whether you are intending to criticize, to praise, to persuade, and so on. Good communication is vital. We hope that by giving you a great deal of pragmatic information in this dictionary, we will encourage you to improve your communication skills.

xiii

Figure 1: Information plate on 'Pragmatics' in the front matter of COBUILD 5

Anglican **a**

anemone /əˈnɛməni/ (anemones) An N-COUNT
anemone is a garden plant with red, purple, or white flowers.

anesthesiologist /æˈnɛsthiːziːɒlɒdʒɪst/ → see anaesthesia.

anesthesiologist /æˈnɛsthiːziːɒlɒdʒɪst/ (anesthesiologists) An N-COUNT
anesthesiologist is a doctor who specializes in giving anaesthetics. [AM]

in BRIT, use **anaesthetist**

anesthetize /əˈnɛsthiːtəɪz/ → see anaesthetize.

anesthetist /əˈnɛsthiːtɪst/ (anesthetists) An N-COUNT
anesthetist is a nurse or other person who gives an anaesthetic to a patient. [AM]

anesthetize /əˈnɛsthiːtəɪz/ → see anaesthetize.

anew /əˈnjuː, AM əˈnuː/ If you do something anew, you do it again, often in a different way from before. [WRITTEN] She's ready to start anew... He began his work anew.

angel /ˈɛɪndʒəl/ (angels) **1** Angels are spiritual beings that some people believe are God's servants in heaven. **2** You can call someone you like very much an **angel** in order to show affection, especially when they have been kind to you or done you a favour. **3** If you describe someone as an **angel**, you mean that they seem to be very kind and good.

angelic /æˈndʒəlɪk/ **1** You can describe someone as **angelic** if they are, or seem to be, very good, kind, and gentle. ...an angelic face... He looked angelic. **2** **Angelic** means like angels or relating to angels. ...angelic choirs.

angelica /æˈndʒəlɪkə/ **Angelica** is the candied stems of the angelica plant which can be used in making cakes or sweets.

anger /ˈæŋɡə/ (angers, angering, angered) **1** **Anger** is the strong emotion that you feel when you think that someone has behaved in an unfair, cruel, or unacceptable way. He cried with anger and frustration... Ellen felt both despair and an-

Figure 2: Random section from COBUILD 5; note the pragmatics data in the extra column (on the right of the articles)

Those added meanings, being descriptions of the ways in which people use language, were based on sound corpus evidence – at the time (in 1995) a digital collection of over two hundred million English words. Since then, the explicit treatment of pragmatics has continued to command increasing dictionary space. In the fully-corpus-driven *Macmillan English Dictionary for Advanced Learners* (MEDAL, Rundell 2002), for example, two entire pages were devoted to pragmatics in the Language Awareness section which is found right in the middle of the central lemma-sign list.

Corpus-based dictionary making in South Africa

In post-1994 South Africa, a brand-new series of dictionaries is being produced for all official languages – Afrikaans, Ndebele, Northern Sotho, South African English, Southern Sotho, Swazi, Tsonga, Tswana, Venda, Xhosa and Zulu. For each of those languages, various corpora have been built, and dictionaries based on the data therein are under compilation. Dictionaries such as COBUILD and MEDAL are held up as an example, and the resulting reference works will be the first ones that will reflect real and authentic use of the South African languages.

This last claim, however, immediately presents a problem. From the above it is clear that there is a direct relationship between the quality of the final products, and the quality of the corpora on which those products are based. Although considerable efforts have been made to create state-of-the-art electronic corpora for all languages involved, technical limitations and difficulties inherent to the specific languages imply that one must be extra careful in interpreting some of the results, as some level of 'transformation' has already taken place during corpus building, with a further transformation taking place during the initial analysis of the corpus data.

Northern Sotho case study

For the purposes of this paper, the focus will be on one language, Northern Sotho. Like all South African corpora, the corpora for Northern Sotho are made up of both digitised and digital material. The former is typically the result of straightforward scanning followed by optical character recognition (OCR), while the latter mainly brings together Internet material and existing electronic files contributed by colleagues, authors, publishers, etc.

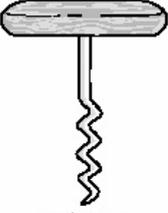
In the Northern Sotho orthography, a distinction is made between the voiceless alveolar fricative 's' and the voiceless palatal fricative 'š'. In order to make sure that any scanned material also makes sense after scanning, it is of paramount importance to ensure that the distinction between these two voiceless fricatives is kept. This means that it is absolutely crucial to 'train' the OCR software so that it also differentiates between the two in all circumstances. (Note that, given the 'š' is not normally part of standard OCR packages, and given the letter 'q' is not used in the Northern Sotho orthography, the OCR software is often trained to recognise and interpret all 'š' as 'q'. A simple search-and-replace then enables one to restore all 'q' to 'š' in the final file.)

True frequencies of use

However, despite rigorous training of the OCR software, the 's' and 'š' are all too often still mixed up. Occurrence frequencies for forms such as 'ntshe' versus 'ntšhe', 'se' versus 'še', 'seba' versus 'šeba', etc. are therefore not fully reliable, which is a problem in corpus-based

lexicography. It is a problem on the macrostructural level of the dictionary, because one cannot indicate the true frequency of use. Dividing the lexicon in 'frequency bands' (by means of lemma signs in colour, various numbers of stars or diamonds accompanying the top-frequent lemmas, etc.) is fast becoming a standard feature in modern dictionaries. See for example Figure 2, illustrating the filled / hollow diamonds as used in COBUILD, or Figures 3 and 4 illustrating star ratings as found in MEDAL and in a Northern Sotho – English dictionary (De Schryver 2007).

cork-screw¹ /'kɔrk,skru/ noun [C] a tool used for pulling the corks out of wine bottles



cork-screw² /'kɔrk,skru/ verb [I] to move in a SPIRAL (=a curve that curls upward)

cork-screw³ /'kɔrk,skru/ adj in a SPIRAL shape, like a corkscrew

cor-mo-rant /'kɔrmərənt/ noun [C] a large dark-colored bird with a long neck that lives near the ocean and eats fish

corn /kɔrn/ noun ★★★

1 [U] a tall plant with large yellow seeds on a cob (=thick piece of stem). *Br E usually maize. 1a.* the seeds of a corn plant that are cooked as food or fed to animals

2 [C] a small piece of hard skin on your foot which is painful

corn-ball /'kɔrn,bɔl/ adj [only before noun] *Am E informal* CORNY

corn bread noun [U] bread made from CORN

corn-cob /'kɔrn,kɔb/ noun [C] the long hard part at the top of a CORN plant, on which large yellow seeds grow

corn-crake /'kɔrn,kreik/ noun [C] a European bird with brown feathers and a loud cry

corn dog noun [C] *Am E* a HOT DOG covered with CORN-MEAL, usually served on a stick

cor-ne-a /'kɔrnio/ noun [C] *medical* the transparent layer that covers the outside of your eye

cor-ne-al /'kɔrnioʊl/ adj *medical* relating to the cornea

corned beef /'kɔrnd 'bif/ noun [U] cooked BEEF that has been preserved in salt water

cor-ner¹ /'kɔrnər/ noun [C] ★★★

1 where two sides meet	5 difficult situation
2 turn/meeting of roads	6 in boxing/wrestling
3 end of mouth/eye	7 in soccer, etc.
4 small (quiet) area	+ PHRASES

1 the part of something square or RECTANGULAR where two edges meet: *Watch the baby, that table has sharp corners. ♦ at/in the corner* The date is displayed in the corner of the screen. ♦ *I had to park in the far corner of the parking lot. ♦ right-hand/left-hand corner* That's me, in

Figure 3: Random section from MEDAL; note the star rating (printed in red)

ntshe *** noun *N-dass* = there ♦ *Ke tšwa ntshe eupša ga ke a mo hwetša. I have just come from there, but I didn't find him.*

Although ntshe is a noun, it is often used as an adverb.

ntšhe * /ntšhe, ntšhè/ verb *c* NTŠHA

1 = (must) take out ♦ *Kgopela bagwera ba gago gore ba go ntšhe diphošo. Ask your parents to take out the mistakes for you.*

2 = (must) discharge **3** = (must) withdraw

♦ *ga/sa/se (...)* **ntšhe** **1** = not take out **2** = not discharge **3** = not withdraw ♦ *Nka se ntšhe tšhelete go lefela thoto yeo e sego ya hlwa e tlišwa. I will not withdraw money and pay for the furniture before it is delivered.*

ntšhetša /ntšhètšə/ verb + applicative (*el*)

c NTŠHA **1** = take out of; take out for ♦ *Mošemane o ile a ntšhetša nonyana ka ndle ga sehloga gomme ya tšhaba. The boy took the bird out of the nest and it flew away.*

2 = withdraw for ♦ *Ke ka fao ke ilego ka kgopela Amos Ratila go yo re ntšhetša tšhelete yeo. That is why I asked Amos Ratila to withdraw the money for us.*

ntšhi noun *9/10* (*pl. dintšhi*) = fly ♦ *Re apeile mogodu ke ka fao go nago le dintšhi tše dintši. We cooked tripe; that is why there are a lot of flies.*

ntšhitše * /ntšhitšè/ verb + perfect (*ile*)

c NTŠHA **1** = took out ♦ *Yo mongwe wa batsomi o be a ntšhitše sethunya a nyaka go thunya tau yeo. One of the hunters took out his rifle to shoot that lion. 2 = withdrew ♦ *Ge a re o lebelela ka pankeng, a hwetša gore o ntšhitše tšhelete ka moka. When she went to check at the bank, she found that she had withdrawn all the money.**

Figure 4: Random section from the PUKUNTŠU YA SEKOLO; note the star rating

Non-standard orthographies

In order to do this correctly for the African-language dictionaries, one would thus need to proofread an entire corpus manually. Apart from the fact that there are no resources available for this, the fact that the spelling has not been standardised throws an extra spanner in the works.

Indeed, take for instance the form 'setše'* which, taken at face value, appears frequently enough in the various Northern Sotho corpora to warrant inclusion in a dictionary; see Figure 5. A closer inspection of all the concordance lines, however, reveals that in 95% of the cases this should actually have been 'šetše' – the result of either scanning errors, or spelling errors in

the original source material. For the remaining 5%, the authors clearly meant to write 'tshetše'. To complicate matters further, there is also a 'tšhetše', where the first 'š' might also be the result of various errors.

N	Concordance	t	T	Word No	File
85	ga Magaga le Mohlago se thomile go biloga. Yeo e be e setše e bonwa le ke sefofu. Iri ya bobedi ge e betha, ke			33,950	\pelotsed.txt
86	makgaebe a lebile kua motseng ka ge sothwane e be e setše e nona 'mme baeti ba nyaka go apeelwa. Motho o			57,105	othowam.txt
87	pelo. Ka nako yeo batho ba be ba robetše ka gore e be e setše e le bošego. Ka mehla kgwedi ka kgwedi lapeng le			69,450	\sesotho5.txt
88	aganywego di šwanetše go laolwa gotee. Afrika Borwa e setše e hwetše meetše gotšwa Lesotho. Bjale ka ge re n			901	kamanole.txt
89	a silafatše tšeo. Panka ya lefase le IMF Afrika Borwa e setše e kgokagane le mokgahlo wa boditšhaba wo matiol			399	kamanole.txt
90	go ngwadiša tšhomišo ya meetse ka lefelelong la gago e setše e tšilw naa. Diforomo tša go ngwadiša di humaneg			2,761	\hlahloya.txt
91	ka mmmino, o bitswa Mello, gomme bjalo tlotlo ya gagwe e setše e didikong tša botagwa - ke ka bjona botagwa bjoo			19,010	\nnetefel.txt
92	e a dulelwa. Banna ge ba bothane kgorong kgoši ke ge e setše e tšo bona ka a yona mahlo tšeo e di tsoseditšwe			12,254	\sekolosa.txt
93	re re ka be re e lelekiše. Ba bangwe ba gana, ba re ge e setše e arogile tsela, e be e tla re šia, ka baka leo re be			17,011	molomats.txt
94	lešupi leo le re le mmadi O tseba koloi yeo, ka gobane e setše e boletšwe mo polelong yeo. Ka tsela yeo tshwant			1,820	\thutadin.txt
95	Mo a lego gona o swanetše go ba a fela pelo ka gore e setše e le sebaka seo se hlalefilego mola ke mo tlogelag			34,443	\dikeledi.txt
96	a thari e swana pheko o filwe: Marapo a dithuto Kubu e setše e a wasantše; A ile kua le kua a re nešetša hlalag			18,660	~1\naledi.txt
97	selepukgomo. Ka yona nako yeo o hwetša bjalo kwete e setše e bo meletša mare. Nka se ke ka re ka lebaka la g			31,319	\pelotsed.txt
98	ba lebatša le tholo yela ba bego ba e rwele ka leako - e setše kua ba. go arogana ntshe ge ba hlabelwa mokgoši			107,347	\kgorongy.txt
99	na re šetše re tswalela lebatli la phapoši ya rena. Nako e setše e ya iring ya lesome. Ra tsena mapaing, ka morag			17,573	\lengwal2.txt
100	mogomo wa yona nameng yeo ka tshwanelo nkabego e setše e bodile. O ile a botšišša batho ba kgauswi gomm			20,265	medupiya.txt
101	sing. Mahlatse, ge re fihla, ra hwetša thekisi ye nngwe e setše e tšala. A namela Refilwe, a se lebale go nthatha k			19,980	\lengwal2.txt
102	setu megokgo ya theoga sefahlegong sa gagwe a sa e setše. O be a sa le a lla la mafelelo kgale e sa le mošem			48,472	\nnetefel.txt
103	koloi yc mpshampsha yela! O ka se hlwe o sa e tseba, e setše e fetogile kotikoti. Thercoo, e be e le ga e sa le ko			29,693	\okolobil.txt
104	fela ba makatšwa ke ge go thwe ba tile polokong. tšeo e setše e le dijo tša ka tša ka mehla. Ka boiketlo ka batam			24,731	\lengwal2.txt
105	a Jerusalem a ya Ntlonggethwa, a lebelediša tšohle. E setše e le lebaka la mantšiboar a tšwa a ya Bethania a n			33,763	\bibmpsha.txt
106	otse. Mongwalelo Matseno Bjalo ka ge kgopolo yeo e setše e hlalošitšwe ge go hlalošwa dikgopolo, mo go yo l			32,568	~1\mjphd.txt
107	o lokeišeneng ba bona lori ya boHarry Kobue le yona e setše e gopaletše mo tseleng kua pele. Bjale ya ba "se			35,165	molomats.txt
108	o e bego e le sa tlhahlo ya barutiši. Go tloga ka 1968 go setše go agilwe dikolo tše phagamego tše seswai mo na			10,222	~1\phupu.txt
109	re Serotologane ke wa gagwe o mo itia makopo. 53 Go setše go fedile dibeke tše tharo morago ga gore Chabala			20,553	\1serotolo.txt
110	nyane bjo a bego a ipoloketše bjona. Mahlatse go be go setše fela dikgwedi tše pedi, gore ba thome ka dithahlob			37,021	\tshelhan.txt
111	labana. A re go Setswatswa: "Re sa tlo le tšaiša, bjale go setše fela ge re le amoga dikgomo le mabele a lena ka			37,839	mogopole.txt
112	mo ya go tšenela phadišano, tobetša mo. Gona bjale go setše dikgwedi tše hlano pele ga tšatšikgwedi ya go tsw			371	\ling0507.txt
113	g bja dimilione tše 268 tša diranta mo nakong ya bjale go setše go thomišitšwe ka tšona, goba di feditšwe, ebile di			2,988	gotswaka.txt
114	helo - kiletšano, Pelo di fišagetše bophelo bja bohle. Go setše matšhalaka bosetšhošwa ke mmutla, Gwa sala ma			1,594	eokgobo.txt
115	leleng e a mpolaya! Banna e ntsemile hlogo ka dinala go setše fela ge e nkgola molala. Seripa sa banna se ile s			20,816	mogopole.txt
116	šu tšeo di ka ba di šetše di momilwe ke mohlwa ka ge go setše go fetile dibeke tše mmalwana. Letswa o ile ge a l			4,373	etshwang.txt
117	tshehlo e tala yeo e sa hlabego. Gape o tsebe gore ge go setše go sele ka maribaneng ka fao e le go ka bogweng			31,220	\nsdigana.txt

Figure 5: The form 'setše'* in a Northern Sotho corpus; here queried with *WordSmith Tools* (Scott 2007)

The inclusion of zero-frequency words

Even entirely trivial cases can lead to radically wrong conclusions. Existing corpus-based dictionaries for Northern Sotho will for example list both 'šoma' / 'šome' and 'soma' / 'some' as different verbal forms (as in the POPULAR, by Kriel, Prinsloo & Sathekge 1997). All four are frequent enough to be included in even a small dictionary. However, reading through all the corpus examples indicates that simply all cases of 'soma' / 'some' should have been 'šoma' / 'šome'. In addition to 'the machine' (i.e. the OCR) which transformed the original, this is a clear case where 'the human user', probably not knowing how to get the 'š' out of a keyboard, simply opted for the erroneous base form without the diacritic throughout. The human users thereby transformed one verb into another, and an inattentive dictionary compiler might end up treating a verb with a zero occurrence!

Collective evidence, or another transformation?

All human users together, however, thanks to the collective evidence in a corpus, may also help in pinpointing the most common spelling for certain words and/or confirm what should

be considered 'the standard' when dealing with an unstable orthography. As such the spelling 'tshepa' is for example one point five times more frequent than 'tshepha', 'bantši' four times more frequent than 'bantšhi', or 'bontši' nine times more frequent than 'bontšhi'. In all these cases, the form without aspiration is clearly the one that needs to be considered as the 'correct' form and thus the one that needs to be included in the dictionary. On the other hand, with this approach one still runs the risk to have factored out dialectal variation in pronunciation, another transformation.

Collective evidence, and dictionary-internal cross-references

Nonetheless, where the difference in frequency between different corpus forms is at least a factor ten, one may safely assume that the more frequently attested corpus form is also 'the norm'. This is for example the case for 'yeo' which is twelve times more frequent than 'eo', 'gešo' which is thirteen times more frequent than 'gešu', or 'kgauswi' which is fourteen times more frequent than 'kgaufsi'. Given that the lesser-used forms remain very frequently used, the lexicographer will do well to cross-refer the lower-frequency forms to the higher-frequency forms, with in each case the reference marker text 'Correct spelling ='. See in this regard Figure 6, which shows the preview of the creation of the articles for 'eo' and 'yeo' in the dictionary compilation environment *TshwaneLex* (Joffe et al. 2007).

eo *** Correct spelling = yeo

Article has cross-references to >>

yeo *** /yêô/ demonstrative pos. II < ye² 1 cl. 9 ► that (one) Karata
 yeo ke ya go reka diaparo. • That card is for buying clothes. 2 cl. 4
 ► those (ones) O iša kae mekotla yeo? • Where are you taking
 those bags?

Note: This demonstrative indicates that the speaker and the addressee are relatively far apart, and that the object/objects referred to, is/are closer to the addressee, but still some distance away from her/him.

Figure 6: Creating the linked articles 'eo' and 'yeo' in the dictionary compilation environment *TshwaneLex*

Collective evidence, and demoting homonyms

In an English dictionary one may for example find under one of the senses of the determiner 'my' the pragmatic information that when 'my' precedes a (pet) name or otherwise kind designation, its function is to show affection. This is done in COBUILD, as can be seen from Figure 7.

my /maɪ/ ◆◆◆

My is the first person singular possessive determiner.

1 A speaker or writer uses **my** to indicate that something belongs or relates to himself or herself.
 I invited him back to my flat for a coffee... John's my best friend. 2 In conversations or in letters, **my** is used in front of a word like 'dear' or 'darling' to show affection. Yes, of course, my darling. 3 **My** is used in phrases such as 'My God' and 'My goodness' to express surprise or shock. [SPOKEN]
 My God, I've never seen you so nervous... My goodness, Tim, you have changed!

Figure 7: The article for 'my' in COBUILD; note the pragmatic information for senses 2 and 3

In Northern Sotho, 'my' is rendered as 'of mine', being a possessive concord in agreement with the noun, followed by the possessive pronoun of the first person singular, thus '[PC +] ka'.

When looking at the various corpora for Northern Sotho one will unfortunately find that forms such as 'a ka', 'ba ka', 'bja ka', 'la ka', 'wa ka' or 'ya ka', which follow the noun they refer to, are often misspelled as single words. Given the high occurrence of these misspelled forms, lexicographers have erroneously lemmatised such forms as single words in their dictionaries in the past, as can for instance be seen in Figure 8, taken from Kriel, Van Wyk & Makopo (1989).

lahlè, id. LH: weggooi.
 lai¹, id. LL: alles/almal vernietig/opeet.
 lai², id. HH: glinster, blink, flits.
 laiki, snw. leenw. kl 9, HHL: laai(tjie).
 laila, ww. LLL (kous. laidiša): oplek, verteer (soos vuur).
 laiša, ww. leenw. HHL: (op)laai.
 laka¹, bsk. kl 5 + bes. st. HL: myne, van my.
 laka², ww. LL: beheer, toesig hou oor, kontroleer, administreer.
 lakaila, ww. LLLL (kous. lakaidiša): oplek, verteer, verwoes, verwyder.
 lakalèla, ww. LLLL (kous. lakalètsa): besorgd/bekommerd wees oor, kla oor; go lakalètsa katlègò, om voorspoed toe te wens.

Figure 8: Random section from the PUKUNTŠU WOORDEBOEK; note that 'laka¹' should be spelled 'la ka', 'laka²' is thus not really a homonym

A lexicographer analysing Northern Sotho corpora when treating 'my' in a bilingual Northern Sotho – English dictionary, will thus need to (a) point out the correct spelling in a Usage Note, and (b) point out, for example by means of an authentic (i.e. corpus-based) example, that the (pragmatic) use whereby 'my' is placed before the noun in English, needs to follow the noun in Northern Sotho. See in this regard Figure 9 (for one instance in the Northern Sotho to English side), which may be contrasted with the information shown in Figure 10 (where one sees the 'grammatical construction' in the English to Northern Sotho side).

baka¹ ** verb ► **cause** Letšhollo le ka baka phepompe. • *Diarrhoea can cause malnutrition.*
baka² ** verb ► **fight over something** Barwa ba kgoši ba a lwa ba baka bogoši bja tatagobona. • *The chief's sons are fighting over their father's chieftainship.*
baka³ ** See **lebaka**
baka⁴ Correct spelling = **ba ka**
ba ka *** possessive construction cl. 2 ► **my; of mine** Bana ba ka ba babedi ba tsena yunibesithi. • *My two children are at university.*

Note: This construction consists of the possessive concord 'ba' of class 2, and the possessive pronoun 'ka' of the first person singular. It is often misspelled as 'baka'.

Figure 9: Creating the article 'ba ka' (as well as making provision for the misspelled 'baka⁴') in the dictionary compilation environment *TshwaneLex*

my *** pronoun ► [PC +] ka I fought with my wife because of him. • *Ke lwele le mosadi wa ka ka lebaka la gagwe.* || I hope my application will be successful. • *Ke holofela gore kgopelo ya ka e tlo atlega.* || Today is my birthday. • *Lehono ke letšatši la ka la matswalo.*
 • (people) of my place/family/homestead ► **bešo** I sent my (family's) people to find me a woman to marry. • *Ke romile batho bešo go ya go nnyalela mosadi.*

Figure 10: Creating the article 'my' in the dictionary compilation environment *TshwaneLex*

Conclusion

The above examples show why it is thus actually not even possible to 'clean up' and to 'read through' an entire (African-language) corpus before starting to use it to derive linguistic products from. The errors and inconsistencies are exactly needed to be able to formulate and derive rules from. Transformations and interpretations have to be performed on the fly with, for dictionary making, corpus-based lexicographic pragmatics one of the logical outcomes.

Dictionary abbreviations

COBUILD 2 = Sinclair 1995

COBUILD 5 = Sinclair 2006

MEDAL = Rundell 2002

POPULAR = Kriel, Prinsloo & Sathekge 1997

PUKUNTŠU WOORDEBOEK = Kriel, Van Wyk & Makopo 1989

PUKUNTŠU YA SEKOLO = De Schryver 2007

References

- De Schryver, Gilles-Maurice.** 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English / Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford.* Cape Town: Oxford University Press Southern Africa.
- Joffe, David et al.** 2007. TshwaneLex Suite. URL = <<http://tshwanedje.com/tshwanelex/>>.
- Kriel, Theunis J., Daniël J. Prinsloo and Bethuel P. Sathekge.** 1997. *Popular Northern Sotho Dictionary, Northern Sotho – English, English – Northern Sotho.* Cape Town: Pharos.
- Kriel, Theunis J., Egidius B. van Wyk and Staupitz A. Makopo.** 1989. *Pukuntšu woordeboek, Noord-Sotho – Afrikaans, Afrikaans – Noord-Sotho.* Pretoria: J.L. van Schaik.
- Rundell, Michael.** 2002. *Macmillan English Dictionary for Advanced Learners.* Oxford: Bloomsbury Publishing Plc.
- Scott, Mike.** 2007. *WordSmith Tools.* URL = <<http://www.lexically.net/wordsmith/>>.
- Sinclair, John M.** 1995. *Collins COBUILD English Dictionary.* London: HarperCollins Publishers.
- Sinclair, John M.** 2006. *Collins COBUILD English Dictionary.* London: HarperCollins Publishers.