

Speech corpora are a valuable tool for studying language in a more naturalistic way and for looking at statistical distributions of patterns and constructions, allowing linguists to judge not only if something is grammatical, but also how common and in what contexts it is used. While extensive speech and text corpora are available for major European languages and increasingly for national languages of Asia, it is much rarer to find corpora for local languages, even those with major speaking populations. In this paper we report on the development of a text and speech corpus for Sundanese (the fourth most widely spoken language in Indonesia).

The Sundanese corpus consists mainly of written language of a rich variety of genres, including more formal styles, from books, papers, articles, speeches, stories, and news items, as well as more casual styles, including song lyrics, websites, blogs, and postings to Facebook. Especially some of the online sources, while “written”, encompass very informal styles. The corpus includes some spoken language as well and this will be further developed as more transcriptions become available through a related project.

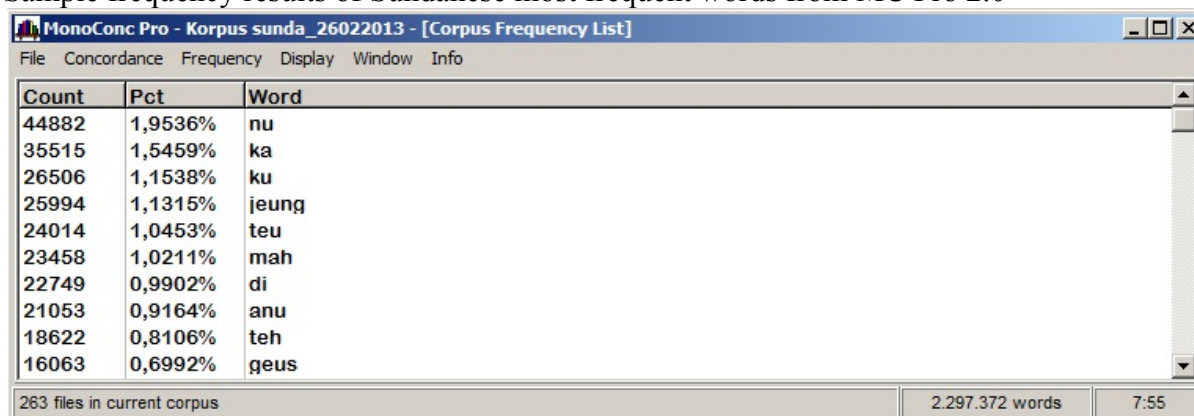
The corpus now comprises around 2.3 million words. While small compared to the well established language corpora of English and other Western European languages, it is an excellent start and can serve as a valuable tool for research, dictionary building etc. The advantage of this corpus in comparison to parallel text retrieval via Google is that the outcomes are much ‘cleaner’, avoiding the extensive noise in Google searches and the need to verify each and every hit.

The software used to develop the corpus is Monoconc Pro 2.0. (<http://www.athel.com/mp.html>), a window-based concordancing program that provides the core functions that corpus linguists typically use. To get started, either text (.txt.) files or material from a URL are loaded into the software, these files together constitute the corpus, across which words and phrases can be searched. The software displays the exact match of the words/phrase plus the context sentences in which they occur. It can also calculate frequency. Search results can be sorted by keyword, allowing similar contexts to be naturally grouped together.

We also present several sample queries to show how the software works and illustrate what kinds of questions can be investigated. The frequency data, for instance, show that, the ten most commonly used words in Sundanese are all functional words, two of which are pragmatic particles, i.e. *mah* [focus] and *téh* [topic] (figure 1). This may suggest the relative importance of such elements in this language. A similar picture is also observed in English corpora (2). Choice of pronoun is a topic of interest. The presence of speech levels makes it clear that, to be deemed polite, a Sundanese speaker should use proper pronouns in appropriate contexts. This is manifested in the corpus, in which certain pronouns occur mainly in certain types of texts. The coarse variant of first person pronoun *déwék* occurs exclusively in narratives (3), while the colloquial counterpart *kuring* occurs in more varied contexts (4), with of course a higher frequency than *déwék*.

Through these examples, we will highlight how corpora can be used to complement other data collection methods and show how straightforward it is to construct a corpus. We hope that this illustration will encourage others to construct corpora of other local languages of Indonesia.

(1) Sample frequency results of Sundanese most frequent words from MC Pro 2.0

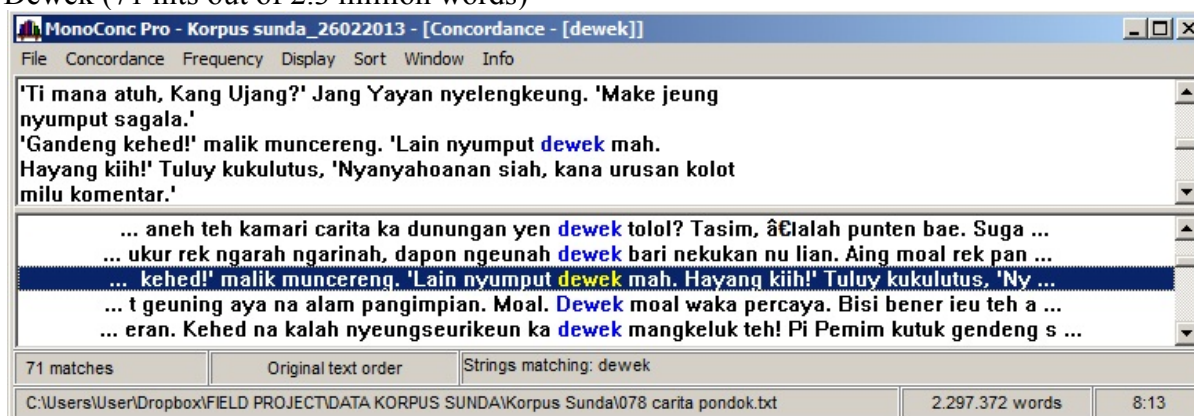


Count	Pct	Word
44882	1,9536%	nu
35515	1,5459%	ka
26506	1,1538%	ku
25994	1,1315%	jeung
24014	1,0453%	teu
23458	1,0211%	mah
22749	0,9902%	di
21053	0,9164%	anu
18622	0,8106%	teh
16063	0,6992%	geus

263 files in current corpus 2.297.372 words 7:55

(2) English top ten most commonly used words are *the, of, to, and, a, in, is, it, you, that* (<http://www.world-english.org/english500.htm>)

(3) Déwék (71 hits out of 2.3 million words)

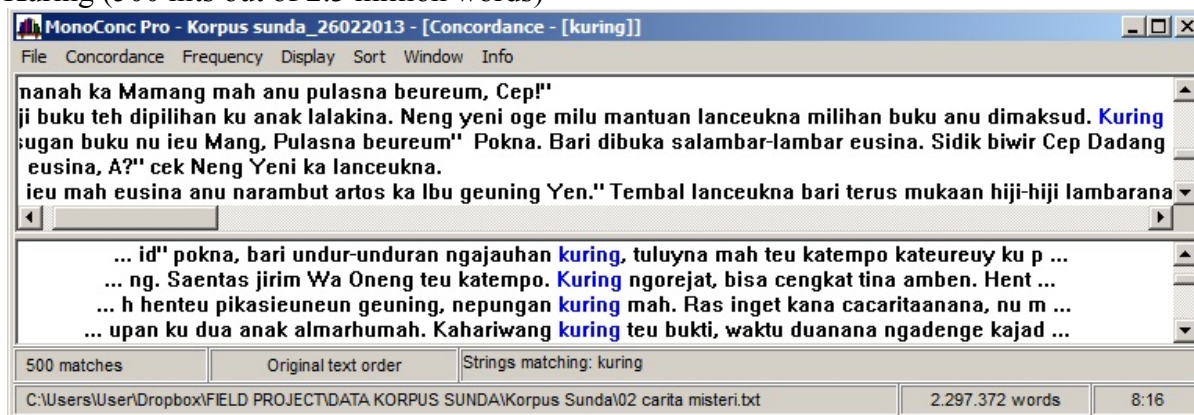


'Ti mana atuh, Kang Ujang?' Jang Yayan nyelengkeung. 'Make jeung nyumput sagala.'
'Gandeng kehed!' malik muncereng. 'Lain nyumput dewek mah. Hayang kiih! Tuluy kukulutus, 'Nyanyahoanan siah, kana urusan kolot milu komentar.'

... aneh teh kamari carita ka dunungan yen dewek tolol? Tasim, â€œlalah punten bae. Suga ...
... ukur rek ngarah ngarinah, dapon ngeunah dewek bari nekukan nu lian. Aing moal rek pan ...
... kehed! malik muncereng. 'Lain nyumput dewek mah. Hayang kiih! Tuluy kukulutus, 'Ny ...
... t geuning aya na alam pangimpian. Moal. Dewek moal waka percaya. Bisi bener ieu teh a ...
... eran. Kehed na kalah nyeungseurikeun ka dewek mangkeluk teh! Pi Pemim kutuk gendeng s ...

71 matches Original text order Strings matching: dewek
C:\Users\User\Dropbox\FIELD PROJECT\DATA KORPUS SUNDA\Korpus Sunda\078 carita pondok.txt 2.297.372 words 8:13

(4) Kuring (500 hits out of 2.3 million words)



nanah ka Mamang mah anu pulasna beureum, Cep!"
ji buku teh dipilihan ku anak lalakina. Neng yeni oge milu mantuan lanceukna milihan buku anu dimaksud. Kuring
sugan buku nu ieu Mang, Pulasna beureum" Pokna. Bari dibuka salambar-lambar eusina. Sidik biwir Cep Dadang
eusina, A?" cek Neng Yeni ka lanceukna.
ieu mah eusina anu narambut artos ka Ibu geuning Yen." Tembal lanceukna bari terus mukaan hiji-hiji lambarana

... id" pokna, bari undur-unduran ngajauhan kuring, tuluyna mah teu katempo kateureuy ku p ...
... ng. Saentas jirim Wa Oneng teu katempo. Kuring ngorejat, bisa cengkat tina amben. Hent ...
... h henteu pikasieuneun geuning, nepungan kuring mah. Ras inget kana cacaritaanana, nu m ...
... upan ku dua anak almarhumah. Kahariwang kuring teu bukti, waktu duanana ngadenge kajad ...

500 matches Original text order Strings matching: kuring
C:\Users\User\Dropbox\FIELD PROJECT\DATA KORPUS SUNDA\Korpus Sunda\02 carita misteri.txt 2.297.372 words 8:16