38 Managing Synchronic Corpus Data with the British National Corpus (BNC)

- Class: Linguistics and Data Science
- Presenter: Sumin Park
- Date: 2023-10-30

Table of content

Section 1. Introduction

Data

British National Corpus World XML edition (BNC)

- Consists of 4,049 files with 10 million words from spoken and 90 million words from written data
- Represents British English of the 1980s

(HINT: YOU CAN PROVIDE WITH MORE INFORMATION ABOUT BNC)

Research topic

English dative alternation between (i) a ditransitive construction and (ii) the often-available prepositional dative to *to*.

Examples:

1. Captain Picard **gave** [Commander Data] [a new phaser]. (ditransitive)
2. Captain Picard **gave** [a new phaser] *to* [Commander Data]. (prepositional)

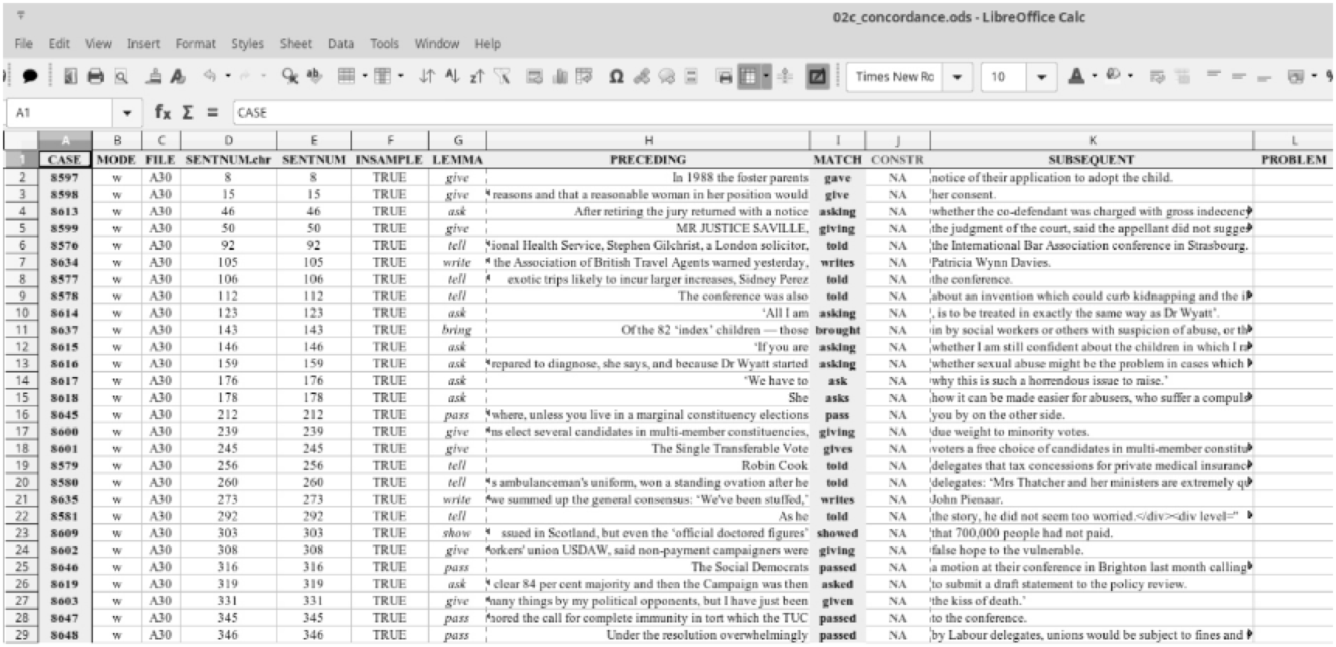
A corpus-linguistic analysis begins from...

Concordance display showing instances of each construction in context

(HINT: YOU MAY WANT TO ADD WHAT A CONCORDANCE DISPLAY IS HERE)

Figure 38.3

Appendix


02c_concordance.ods - LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Times New Ro
10
A1
CASE

	A	B	C	D	E	F	G	H	I	J	K	L
1	CASE	MODE	FILE	SENTNUM.chr	SENTNUM	INSAMPLE	LEMMA	PRECEDING	MATCH	CONSTR	SUBSEQUENT	PROBLEM
2	8597	w	A30	8	8	TRUE	*give*	In 1988 the foster parents	**gave**	NA	notice of their application to adopt the child.	
3	8598	w	A30	15	15	TRUE	*give*	reasons and that a reasonable woman in her position would	**give**	NA	her consent.	
4	8613	w	A30	46	46	TRUE	*ask*	After retiring the jury returned with a notice	**asking**	NA	whether the co-defendant was charged with gross indecenc	
5	8599	w	A30	50	50	TRUE	*give*	MR JUSTICE SAVILLE,	**giving**	NA	the judgment of the court, said the appellant did not sugge	
6	8576	w	A30	92	92	TRUE	*tell*	ional Health Service, Stephen Gilchrist, a London solicitor,	**told**	NA	the International Bar Association conference in Strasbourg.	
7	8634	w	A30	105	105	TRUE	*write*	the Association of British Travel Agents warned yesterday,	**writes**	NA	Patricia Wynn Davies.	
8	8577	w	A30	106	106	TRUE	*tell*	exotic trips likely to incur larger increases, Sidney Perez	**told**	NA	the conference.	
9	8578	w	A30	112	112	TRUE	*tell*	The conference was also	**told**	NA	about an invention which could curb kidnapping and the i	
10	8614	w	A30	123	123	TRUE	*ask*	'All I am	**asking**	NA	, is to be treated in exactly the same way as Dr Wyatt'.	
11	8637	w	A30	143	143	TRUE	*bring*	Of the 82 'index' children — those	**brought**	NA	in by social workers or others with suspicion of abuse, or th	
12	8615	w	A30	146	146	TRUE	*ask*	'If you are	**asking**	NA	whether I am still confident about the children in which I r	
13	8616	w	A30	159	159	TRUE	*ask*	repared to diagnose, she says, and because Dr Wyatt started	**asking**	NA	whether sexual abuse might be the problem in cases which	
14	8617	w	A30	176	176	TRUE	*ask*	'We have to	**ask**	NA	why this is such a horrendous issue to raise.'	
15	8618	w	A30	178	178	TRUE	*ask*	She	**asks**	NA	how it can be made easier for abusers, who suffer a compuls	
16	8645	w	A30	212	212	TRUE	*pass*	where, unless you live in a marginal constituency elections	**pass**	NA	you by on the other side.	
17	8600	w	A30	239	239	TRUE	*give*	ns elect several candidates in multi-member constituencies,	**giving**	NA	due weight to minority votes.	
18	8601	w	A30	245	245	TRUE	*give*	The Single Transferable Vote	**gives**	NA	voters a free choice of candidates in multi-member constitu	
19	8579	w	A30	256	256	TRUE	*tell*	Robin Cook	**told**	NA	delegates that tax concessions for private medical insuranc	
20	8580	w	A30	260	260	TRUE	*tell*	s ambulanceman's uniform, won a standing ovation after he	**told**	NA	delegates: 'Mrs Thatcher and her ministers are extremely q	
21	8635	w	A30	273	273	TRUE	*write*	we summed up the general consensus: 'We've been stuffed,'	**writes**	NA	John Pienaar.	
22	8581	w	A30	292	292	TRUE	*tell*	As he	**told**	NA	the story, he did not seem too worried.</div><div level="	
23	8609	w	A30	303	303	TRUE	*show*	ssued in Scotland, but even the 'official doctored figures'	**showed**	NA	that 700,000 people had not paid.	
24	8602	w	A30	308	308	TRUE	*give*	orkers' union USDAW, said non-payment campaigners were	**giving**	NA	false hope to the vulnerable.	
25	8646	w	A30	316	316	TRUE	*pass*	The Social Democrats	**passed**	NA	a motion at their conference in Brighton last month calling	
26	8619	w	A30	319	319	TRUE	*ask*	clear 84 per cent majority and then the Campaign was then	**asked**	NA	to submit a draft statement to the policy review.	
27	8603	w	A30	331	331	TRUE	*give*	nany things by my political opponents, but I have just been	**given**	NA	the kiss of death.'	
28	8647	w	A30	345	345	TRUE	*pass*	nored the call for complete immunity in tort which the TUC	**passed**	NA	to the conference.	
29	8648	w	A30	346	346	TRUE	*pass*	Under the resolution overwhelmingly	**passed**	NA	by Labour delegates, unions would be subject to fines and	

Figure 38.3
Concordance display.

Three Steps for corpus-linguistic search process

1. Running a query/search based on much existing annotation.
2. Preparing the concordance lines for annotation.
3. Doing some final cheking and preparatory steps for the following statistical analysis.

(HINT: A PROPER NUMBERING MATTERS)

Step 1: Running a query/search based on much existing annotation

Query/search: Words/lemmas and parts of speech.

Ten verb lemmas selected:

- Four verb lemmas that strongly prefer the ditransitve: *tell*, *give*, *show*, and *ask*.
- Three verb lemmas that strongly prefer the prepositional dative: *bring*, *sell*, and *pass*.
- Three verb lemmas that are relatively neutral with regard to the two constructions: *send*, *lend*, and *write*.

Step 2: Preparing the concordance lines for two kinds of annotation

1. To identify the hits that involve the verbs but not the constructions in question.
2. To annotate each constructional use for the linguistic/contextual variables whose effect on the dative alternation is to be studied.

Step 3: Doing some final cheking and preparatory steps for the following statistical analysis

(HINT: SUCH A STATISTICAL ANALYSIS IS OUT OF TOPIC HERE)

Section 2. Retrieval

First step of data management

Extracting a first version of the concordance lines **from the BNC**.

(HINT: WHAT DOES THE BNC ANNOTATION LOOK LIKE? SEE THE FOLLOWING EXAMPLE)

Example 38.1

Two one-sentence utterances from the BNC World edition, file D8Y.xml.

```xml
<!-- Utterance (speaker id: D8YPS006) -->
<u who="D8YPS006">
    <!-- Sentence (sentence number: 80) -->
    <s n="80">
        <!-- Word "And" (POS tag: CONJ-CJC, Head word: and) -->
        <w c5="CJC" hw="and" pos="CONJ">
            And
        </w>
        <!-- Word "erm" (POS tag: UNC-UNC, Head word: erm) -->
        <w c5="UNC" hw="erm" POS="UNC">
            erm
        </w>
        <!-- Pause -->
        <pause/>
        ...
        <!-- Punctuation "." -->
        <c c5="PUN">
            .
        </c>
    </s>
</u>
```

(HINT: ADD AS MANY COMMENTS AS POSSIBLE TO THE XML FILE ABOVE)

Task

Creating a concordance of the ten above-mentioned verb lemmas from all of the BNC.

(HINT: THE AUTHOR USES THE R LANGUAGE HERE, THEN YOU DON'T NEED TO DETAIL THE ALGORITHM HERE)

Requirement

Each use of the given ten verbs is in its own row, as in Table 38.1

Table 38.1

The case-by-variable format for two matches in one sentence

Preceding	Match	Subsequent

Preceding	Match	Subsequent
Picard	showed	Data a phaser then Data showed it to Riker
Picard showed Data a phaser and then Data	showed	it to Riker

- The sentence *"Picard showed Data a phaser and then Data showed it to Riker"* contains two matches for *showed*, so each of the both is represented as an independent row.
- **Preceding**: the preceding context of each *showed*.
- **Match**: the use of each *showed*.
- **Subsequent**: the following context of each *showed*.

(HINT: THE PRECEDING/SUBSEQUENT CONTEXTS MUST BE PROVIDED FOR A FUTURE STUDY)

Final product

Save the concordance data as a CSV file.

Section 3. Annotation

Second steps to data manangement

1. Preparing to weed out false positives.
2. Adding annotation with regard to the variables that might affect the dative alternation to the true positives.

(HINT: IF YOU ARE NOT FAMILIAR TO TRUE/FALSE POSITIVE/NEGATIVES, RESEARCH AND EXPLAIN HERE)

Tasks

1. Adding additional columns to the data

   - **Construction**: a column that will contain the labels *ditrans*/*prepdat*/*other*.
   - **Problem**: a column that will contain the letter of the column that contain something problematic.
     - Don't change the color of the problematic cell!

2. Selecting a sample of concordance lines

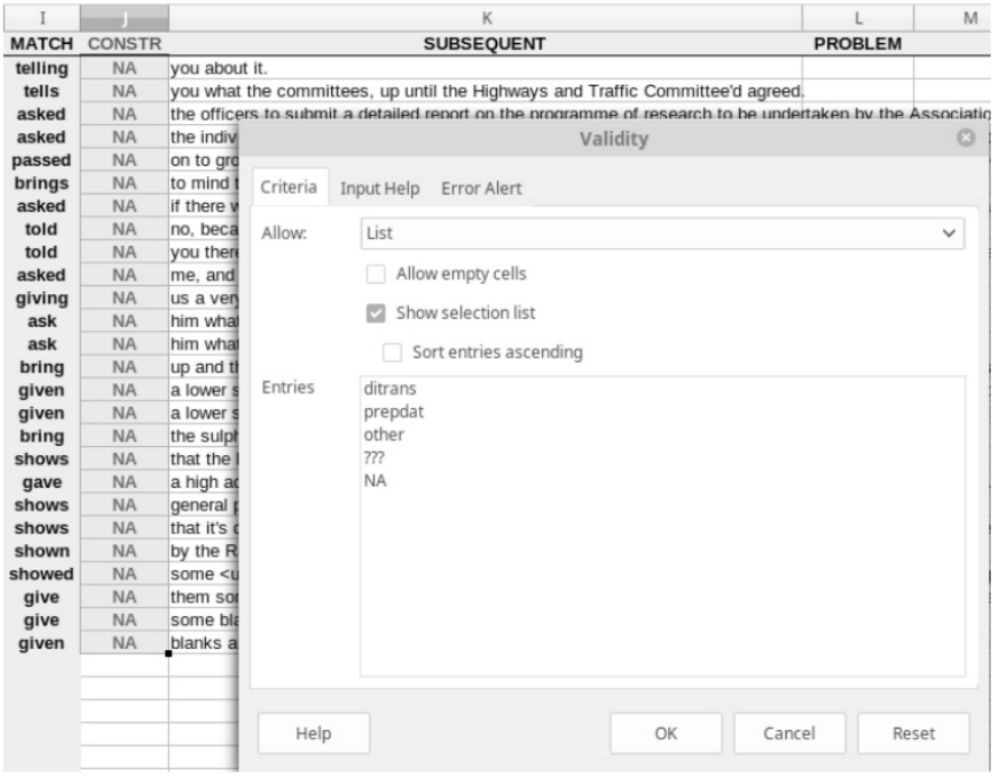
   - Don't draw random sample of lines.
   - Do draw random sample of files. (HINT: WHY?)
     - Set a random-number seed for replicability!

3. Annotating actually (i) whether concordance lines are the right construction(s) and (ii) what their characteristics are that might have affected the speaker's choice

   - Use a spreadsheet software such as LibreOffice Calc.

Figure 38.1

The use of the [Data: Validity] function to guide and constrain data entry.



I	J	K	L	M
MATCH	CONSTR	SUBSEQUENT	PROBLEM	
telling	NA	you about it.		
tells	NA	you what the committees, up until the Highways and Traffic Committee'd agreed.		
asked	NA	the officers to submit a detailed report on the programme of research to be undertaken by the Associatio		
asked	NA	the indiv		
passed	NA	on to gro		
brings	NA	to mind t		
asked	NA	if there w		
told	NA	no, beca		
told	NA	you there		
asked	NA	me, and		
giving	NA	us a very		
ask	NA	him what		
ask	NA	him what		
bring	NA	up and t		
given	NA	a lower s		
given	NA	a lower s		
bring	NA	the sulph		
shows	NA	that the		
gave	NA	a high ac		
shows	NA	general p		
shows	NA	that it's c		
shown	NA	by the R		
showed	NA	some <u		
give	NA	them so		
give	NA	some bla		
given	NA	blanks a		

Validity

Allow: List

☐ Allow empty cells

☑ Show selection list

☐ Sort entries ascending

Entries

ditrans
prepdat
other
???
NA

Figure 38.1
The use of the Data: Validity function to guide and constrain data entry.

Section 4. Preparation and documentation

Third step to data manangement

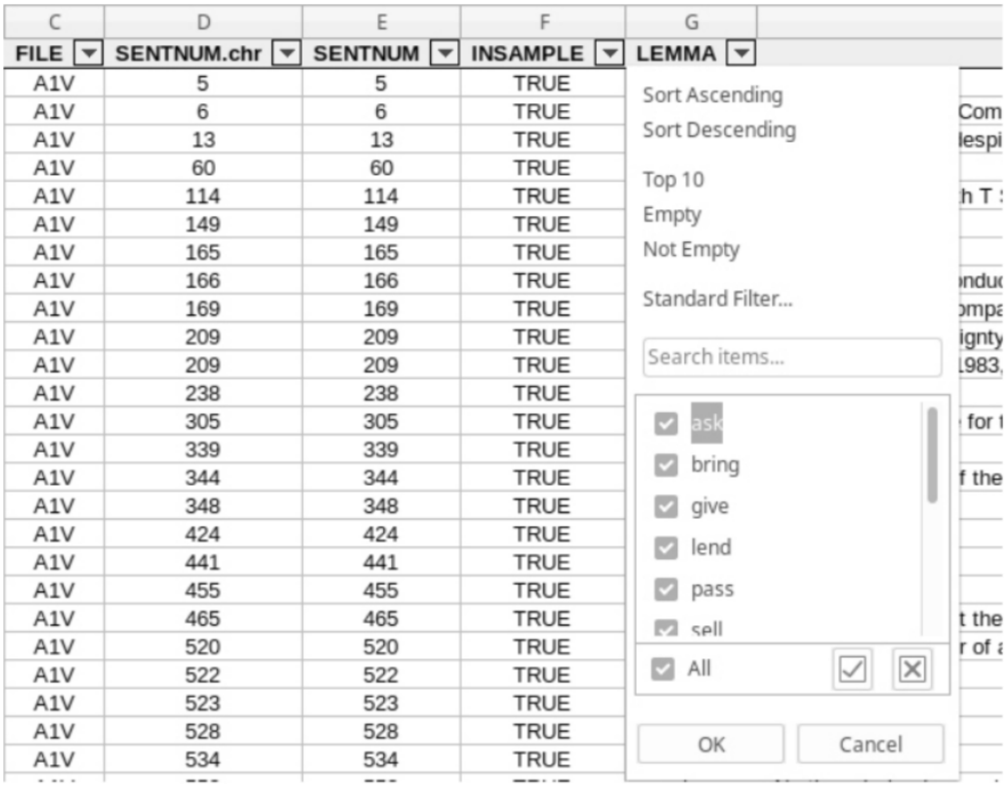
Making sure the data are in good shape for subsequent analysis

Task

Performing a general sanity check of the data as a whole but specifically the annotation that was entered.

Figure 38.2

The use of the [Data: Autofilter] to check data entry.



C	D	E	F	G	
FILE	SENTNUM.chr	SENTNUM	INSAMPLE	LEMMA	
A1V	5	5	TRUE		
A1V	6	6	TRUE		Com
A1V	13	13	TRUE		lespi
A1V	60	60	TRUE		
A1V	114	114	TRUE		h T
A1V	149	149	TRUE		
A1V	165	165	TRUE		
A1V	166	166	TRUE		nduc
A1V	169	169	TRUE		mpa
A1V	209	209	TRUE		ignty
A1V	209	209	TRUE		1983,
A1V	238	238	TRUE		
A1V	305	305	TRUE		for t
A1V	339	339	TRUE		
A1V	344	344	TRUE		f the
A1V	348	348	TRUE		
A1V	424	424	TRUE		
A1V	441	441	TRUE		
A1V	455	455	TRUE		
A1V	465	465	TRUE		t the
A1V	520	520	TRUE		r of a
A1V	522	522	TRUE		
A1V	523	523	TRUE		
A1V	528	528	TRUE		
A1V	534	534	TRUE		

Figure 38.2 The use of Data: Autofilter to check data entry.

Another task

Naming files and documenting workflow across files.

Quiz question

What elements does a concordance display consist of?