Lecture 9: Corpus Linguistics

Ling 1330/2330 Intro to Computational Linguistics
Na-Rae Han, 9/26/2023

Objectives

- Corpus linguistics
- Corpus types
- Key concepts

- Making assumptions: pitfalls
- Linguistic analysis and interpretation: how to draw conclusions

ConditionalFreqDist vs. not-found keys

```
>>> chars = list('colorless green ideas sleep')
>>> chars
    ['c', 'o', 'l', 'o', 'r', 'l', 'e', 's', 's', ' ', 'g', 'r', 'e', 'e', 'n', ' ', 'i',
    'd', 'e', 'a', 's', ' ', 's', 'l', 'e', 'e', 'p']
>>> chars_2grams = list(nltk.bigrams(chars))
>>> chars_2grams
    [('c', 'o'), ('o', 'l'), ('l', 'o'), ('o', 'r'), ('r', 'l'), ('l', 'e'), ('e', 's'),
    ('s', 's'), ('s', ' '), (' ', 'g'), ('g', 'r'), ('r', 'e'), ('e', 'e'), ('e', 'n'),
    ('n', ' '), (' ', 'i'), ('i', 'd'), ('d', 'e'), ('e', 'a'), ('a', 's'), ('s', ' '), ('
    ', 's'), ('s', 'l'), ('l', 'e'), ('e', 'e'), ('e', 'p')]
>>> chars_2gram_cfd = nltk.ConditionalFreqDist(chars_2grams)
>>> chars_2gram_cfd['e']
    FreqDist({'e': 2, 's': 1, 'n': 1, 'a': 1, 'p': 1})
>>> len(chars_2gram_cfd)
    12
>>> chars_2gram_cfd['x']
    FreqDist({})
>>> len(chars_2gram_cfd)
    13
>>> chars_2gram_cfd.keys()
    dict_keys(['c', 'o', 'l', 'r', 'e', 's', ' ', 'g', 'n', 'i', 'd', 'a', 'x'])
>>> chars_2gram_cfd['4']['2']
    0
>>> len(chars_2gram_cfd)
    14
>>> chars_2gram_cfd.keys()
    dict_keys(['c', 'o', 'l', 'r', 'e', 's', ' ', 'g', 'n', 'i', 'd', 'a', 'x', '4'])
```

When a key is not found, NLTK's CFD inserts it as a new key!!

FreqDist does not share this bug.

Be careful: don't trust `len()` of CFD.

So, what does a corpus really look like?

1. They can just look like a bunch of ordinary text files
   - Raw corpora

2. Or they can look a bit more complex, with more bits of information embedded…
   - Raw corpora with html/xml tags
   - Annotated corpora (part of speech, syntactic structures, etc.)

XML format, POS and lemma information

The BNC

```
    <w pos="ADJ" hw="scottish" c5="AJ0">Scottish </w>
    <w pos="SUBST" hw="city" c5="NN2">cities</w>
    <c c5="PUN">.</c>
  </s>
- <s n="136">
    <w pos="SUBST" hw="church" c5="NN2">Churches </w>
    <w pos="PREP" hw="in" c5="PRP">in </w>
    <w pos="ADJ" hw="these" c5="DT0">these </w>
    <w pos="SUBST" hw="area" c5="NN2">areas </w>
    <w pos="ADV" hw="particularly" c5="AV0">particularly </w>
    <w pos="VERB" hw="need" c5="VVB">need </w>
    <w pos="PREP" hw="to" c5="TO0">to </w>
    <w pos="VERB" hw="be" c5="VBI">be </w>
    <w pos="VERB" hw="inform" c5="VVN">informed</w>
    <c c5="PUN">, </c>
    <w pos="ADJ" hw="involved" c5="AJ0">involved </w>
    <w pos="PREP" hw="in" c5="PRP">in </w>
    <w pos="SUBST" hw="community" c5="NN1">community </w>
    <w pos="SUBST" hw="care" c5="NN1">care </w>
    <w pos="CONJ" hw="and" c5="CJC">and </w>
    <w pos="ADJ" hw="supporting" c5="AJ0-VVG">supporting </w>
    <w pos="ADJ" hw="christian" c5="AJ0">Christian </w>
    <w pos="SUBST" hw="worker" c5="NN2">workers </w>
    <w pos="VERB" hw="seek" c5="VVG">seeking </w>
    <w pos="PREP" hw="to" c5="TO0">to </w>
    <w pos="VERB" hw="prevent" c5="VVI">prevent </w>
    <w pos="ADJ" hw="new" c5="AJ0">new </w>
    <w pos="SUBST" hw="hiv" c5="NP0">HIV </w>
    <w pos="SUBST" hw="infection" c5="NN1">infection </w>
    <w pos="PREP" hw="in" c5="PRP">in </w>
    <w pos="SUBST" hw="school" c5="NN2">schools</w>
    <c c5="PUN">.</c>
    <c c5="PUQ">’</c>
  </s>
</p>
```

Syntactically annotated

The Penn Treebank

```
            (PP (IN of)
              (NP
                (QP (CD 118.6) (CD million) )
                (NNS shares) )))))
    (. .) ))
( (S
    (PP (IN In)
      (NP
        (NP (NNS terms) )
        (PP (IN of)
          (NP (NN volume) ))))
    (, ,)
    (NP-SBJ (PRP it) )
    (VP (VBD was)
      (NP-PRD
        (NP (DT an) (JJ inauspicious) (NN beginning) )
        (PP (IN for)
          (NP (NNP November) ))))
    (. .) ))
( (S
    (NP-SBJ
      (NP (NN Yesterday) (POS 's) )
      (NN share) (NN turnover) )
    (VP (VBD was)
      (PP-LOC-PRD
        (ADVP (RB well) )
```

Sentence pair in parallel corpus

```
;;05:127: 저는 그 일을 할 수 있는 한 빨리 하겠습니다 .

(S (NP-SBJ 저/NPN+는/PAU)
   (VP (NP-OBJ-LV 그/DAN
                  일/NNC+을/PCA)
       (VP (NP-ADV (S (NP-SBJ (S (NP-SBJ *pro*)
                                 (VP 하/VV+ㄹ/EAN))
                              (NP 수/NNX))
                         (ADJP 있/VJ+는/EAN))
                   (NP 한/NNX))
           (ADVP 빨리/ADV)
           (VP (LV 하/VV+겠/EPF+습니다/EFN))))
   ./SFN)
```

```
;;05:127: I will do it as fast as possible.

(S (NP-SBJ (PRP I))
   (VP (MD will)
       (VP (VP (VB do)
               (NP-OBJ (PRP it)))
           (ADVP-MNR (ADVP (RB as)
                           (RB fast))
                     (PP (IN as)
                         (ADJP (JJ possible))))))
   (. .))
```

Korean English Treebank Annotations: https://catalog.ldc.upenn.edu/LDC2002T26

Key concepts in corpus linguistics

- Type, token
- Type-token ratio (TTR)
- Frequency
  - Zipf's law
- Concordance
- Collocation
- *n*-grams ("chunks")
- Sparseness problem

Counting words: token, type, TTR

- **Word *token***: each word occurring in a text/corpus
  - Corpora sizes are measured as total number of words (=tokens)
- **Word *type***: unique words
  - Q: Are 'sleep' and 'sleeps' different types or the same type?
  - A: Depends.
    - Occasionally, types are meant as *lemma* types.
    - Typically, inflected and derived words count as different types.
    - Sometimes, even capitalized vs. lowercase words count as 2 types.

← Pay attention to how types are handled in your resource!

- *Hapax legomena* ("hapaxes")
  - Words that occur only once.
  - In natural language corpora, a huge portion will typically be hapaxes.
- **Type-Token Ratio** (**TTR**)

➔ Next slide

Type-token ratio

- **Type-Token Ratio (TTR)**
  - The number of types divided by the number of tokens
  - Often used as an indicator of **lexical density / vocabulary diversity.** (with caveat!)

'Rose is a rose is a rose is a rose'

3/10 = 0.3 TTR

'A rose is a woody perennial flowering plant of the genus Rosa'

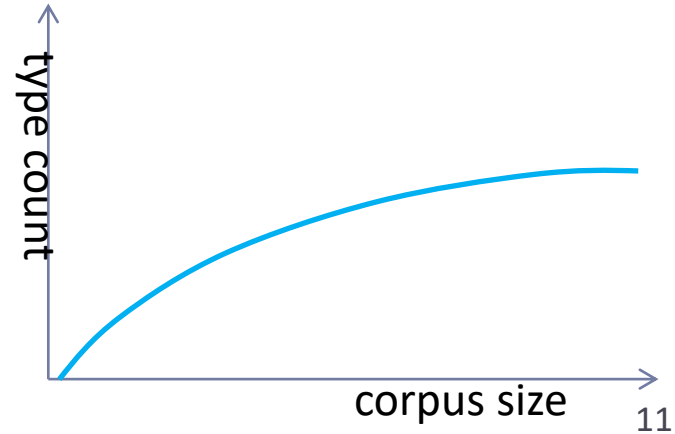
11/12 = 0.916 TTR

Type-token ratio: the caveat

- *Alice in Wonderland*
  - 2,585 types / 35,656 tokens = 0.097 TTR
- *Moby Dick*
  - 17,172 types / 265,010 tokens = 0.080 TTR

← Does this mean Moby Dick has less diverse vocabulary?

- Not necessarily -- the **text sizes** are different.
- **Type #** does not grow linearly with text size. As your text grows larger, fewer and fewer new word types will be encountered.
- TTR comparison is only meaningful for **comparably sized texts**.



Word frequency

- The words in a corpus can be arranged in order of their frequency in that corpus
- Comparing frequency lists across corpora can highlight differences in register and subject field

- Frequency distribution in natural language texts observes common patterns:
  - Word frequencies are not distributed evenly.
  - A small number of words are found in large frequencies
  - Long tails: a large number of words found in small frequencies (tons of hapaxes!)

Example: Tom Sawyer

- Word tokens: 71,370
- Word types: 8,018
- TTR: 0.11
- Top word frequencies: →
- Frequencies of frequencies:

Word frequency	# of word types with the frequency
1	3993
2	1292
3	664
4	410
5	243
…	…
51-100	99
> 100	102

Over 90% of word types occur 10 times or less.

Word	Freq
the	3332
and	2972
a	1775
to	1725
of	1440
was	1161
it	1027
in	906
that	877
he	877
I	783
his	772
you	686
Tom	679

Zipf's Law

- Published in *Human Behavior and the Principle of Least Effort (1949)*
- Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.
  - 1st common word: twice the count of 2nd most common word
  - 50th common word: 3-times the count of 150th common word
- Holds true in natural language corpora!
  - *Tom Sawyer* →

Word	Frequency	Rank
one	172	50
two	104	100
turned	51	200
name	21	400
friends	10	800

Concordances

- **Concordance** lines: shows instances of query words/ phrases found in a corpus; produced by concordance programs
- **KWIC**: "Key Word in Context"
- Lextutor concordancer: https://lextutor.ca/conc/eng/

acters in modern dress and often	**achieve**	**INTERESTING** results. So far as I know, the Comedie
his doesn't detract from its merit as	**an**	**INTERESTING**, if not great, film. The Chronicle's Pa
can look for improvement this year is	**an**	**INTERESTING** one. You'd never guess it from the way
release is to say that it reproduces	**an**	**INTERESTING** and effective Steinberg performance wit
panels decorate a 60-inch long chest.	**An**	**INTERESTING** approach to the bedroom is presented,
er Hotel on Lincoln Park West. It was	**an**	**INTERESTING** fraternisation of ex-convicts, union ra
iscal session of the Legislature with	**an**	**INTERESTING** dilemma. Since the constitution forbids
that most weasel-worded of encomiums "	**an**	**INTERESTING** experiment". There are few things of wh
why this process occurs would provide	**an**	**INTERESTING** separate subject for study. In some are
modern devices in American homes, but	**an**	**INTERESTING** blend of cultures finds us using Japane

Collocation

- **Collocation**: statistical tendency of words to co-occur
- Concordance lines are used by humans to mostly observe; collocation data is processed and compiled by computerized statistical operations
- ex. **collocates** of *shed*:
  - *light*, *tear/s*, *garden*, *jobs*, *blood*, *cents*, *image*, *pounds*, *staff*, *skin*, *clothes*
- Strong collocations indicate:
  - finer-grained semantic compatibility, polysemy
  - association between a lexical item and its frequent grammatical environment
  - lexical collocates of *head*:
    - *SHAKE*, *injuries*, *SHOOT*
    - *state*, *office*, *former*, *department*
  - grammatical collocates of *head*:
    - *of*, *over*, *on*, *back*, *off*

n-grams, chunks, *n*-gram frequency

- Units of word sequences are called
  - ***n-grams*** in computational linguistics
  - ***chunks*** in corpus linguistics circles
  - Certain chunks (*a couple of*, *at the moment*, *all the time*) are as frequent as ordinary, everyday single words such as *possible*, *alone*, *fun*, *expensive*.
- *n*-grams are of interest to computational linguists because they are the backbones of statistical language modeling
- Chunks are of interest in:
  - applied linguistics because they are markers of successful second-language acquisition
  - corpus linguistics because they highlight stylistic and register variation

- *n*-gram frequencies, general vs. written register

Top 3-word chunks, North American English	
1	*I don't know*
2	*a lot of*
3	*you know what*
4	*what do you*
5	*you have to*
6	*I don't think*
7	*I was like*
8	*you want to*
9	*do you have*
10	*I have to*
11	*I want to*
12	*I mean I*
13	*a little bit*
14	*you know I*
15	*one of the*
16	*and I was*

Top 3-word chunks, Written English	
1	*one of the*
2	*out of the*
3	*it was a*
4	*there was a*
5	*the end of*
6	*a lot of*
7	*there was no*
8	*as well as*
9	*end of the*
10	*to be a*
11	*it would be*
12	*in front of*
13	*it was the*
14	*some of the*
15	*I don't know*
16	*on to the*

Data sparseness problem

- In natural language data, frequent phenomena are *very* frequent, while the majority of data points remain relatively rare. (← Zipf's law)
- As your consider larger linguistic context, it compounds the **data sparseness/sparsity problem**.
  - For 1-grams, Norvig's 333K 1-gram data was plenty.
  - Assuming 100K English word types, for bigrams we need 100K**2 = 10,000,000,000 data points.
  - Norvig's bigram list was 250K in size: nowhere near adequate. Even COCA's 1 million does not provide good enough coverage. (cf. Google's original was 315 mil.)
  - Numbers grow exponentially as we consider 3-grams, 4- and 5-grams!

2-grams: Norvig vs. COCA

- count_2w.txt

```
you get          25183570
you getting      430987
you give         3512233
you go           8889243
you going        2100506
you gone         210111
you gonna        416217
you good         441878
you got          4699128
you gotta        668275
you graduate     117698
you grant        103633
you great        450637
you grep         120367
you grew         102321
you grow         398329
you guess        186565
you guessed      295086
you guys         5968988
you had          7305583
you hand         120379
```

Total # of entries:
← ¼ million*
vs.
1 million →

*NOT google's fault!
Norvig only took top
0.1% of 315 million.

- w2_.txt

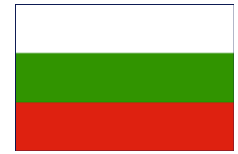
```
39509   you     get
30      you     gets
31      you     gettin
861     you     getting
263     you     girls
24      you     git
5690    you     give
138     you     given
169     you     giving
182     you     glad
46      you     glance
23594   you     go
70      you     god
54      you     goddamn
115     you     goin
9911    you     going
1530    you     gon
262     you     gone
444     you     good
25      you     google
19843   you     got
```

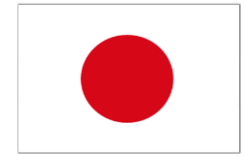
What "you g…" bigrams are missing?

Corpus-based linguistic research

- Create a corpus of Tweets, analyze linguistic variation (based on geography, demographics, etc.)
  - https://languagelog.ldc.upenn.edu/nll/?p=3536
- Process the inaugural speeches of all US presidents, analyze trends in sentence and word length
  - https://languagelog.ldc.upenn.edu/nll/?p=3534
- A corpus of rap music lyrics: word/bigram/trigram frequencies?
  - https://pudding.cool/2017/02/vocabulary/index.html
- Corpora of female and male authors. Any stylistic differences?
- **Corpora of Japanese/Bulgarian English learners. How do their Englishes compare?**

HW 3: Two EFL Corpora





Bulgarian Students

It is time, that our society is dominated by industrialization. The prosperity of a country is based on its enormous industrial corporations that are gradually replacing men with machines. Science is highly developed and controls the economy. From the beginning of school life students are expected to master a huge amount of scientific data. Technology is part of our everyday life.

Children nowadays prefer to play with computers rather than with our parents' wooden toys. But I think that in our modern world which worships science and technology there is still a place for dreams and imagination.

There has always been a place for them in man's life. Even in the darkness of the …

Japanese Students

I agree greatly this topic mainly because I think that English becomes an official language in the not too distant. Now, many people can speak English or study it all over the world, and so more people will be able to speak English. Before the Japanese fall behind other people, we should be able to speak English, therefore, we must study English not only junior high school students or over but also pupils. Japanese education system is changing such a program. In this way, Japan tries to internationalize rapidly. However, I think this way won't suffice for becoming international humans. To becoming international humans, we should study English not only school but also daily life. If we can do it, we are able to master English conversation. It is important for us to master English honorific words. …

Assessing writing quality

- Measurable indicators of writing quality

1. **Syntactic complexity**
   - Long, complex sentences vs. short. simple sentences

   ← Average sentence length, types of syntactic clauses used
2. **Lexical diversity**
   - Diverse vocabulary used vs. small set of words repeatedly used

   ← Type-token ratio (with caveat!) or other measures
3. **Vocabulary level**
   - Common, everyday words vs. sophisticated & technical words

   ← Average word length (common words tend to be shorter)

   ←% of word tokens in top 1K, 2K, 3K most common English words (Google Web 1T n-grams!)

Corpus analysis: beyond numbers

- We have become adept at processing corpora to produce these metrics:
  - How big is a corpus? How many unique words?
    - ← # of tokens, # of types
  - Unigram, bigram, n-gram frequency counts
    - ← Which words and n-grams are frequent

But:
what you get is a whole lot of NUMBERS.

INTERPRETATION of these numbers is what really matters.

Corpus analysis: pitfalls

1. It's too easy to get hyper-focused on the coding part and lose sight of the **actual linguistic phenomena** behind it all.
   ← Make sure to maintain your linguistic motivation.
2. As your corpus gets large, you run the risk of ***flying blind***: it is difficult to keep tabs on what linguistic data you are handling.
   ← Poke your data in **shell**. Take time to understand the data.
   ← Make sure your data object is ***correct***. Do NOT just assume it is.
   ← Scrutinize figures. Are they too small/large? Maybe something's wrong.
3. Attaching a **linguistically valid interpretation** to numbers is not at all trivial.
   ← Careful when **drawing conclusions**.
   ← Make sure to explore **all factors** that might be affecting numbers.

Homework 3

- This homework assignment is equally about Python coding **AND corpus analysis**.
- That means, calculating the **correct numbers** is no longer enough.
- You should take care to understand your corpus data and offer up **well-considered and valid analysis** of the data points.

Wrap-up

- Homework 3 is due on THU
  - Larger at 60 points
- Next class:
  - HW3 review
  - Classifying documents