Corpus Linguistics

A Simple Introduction

Dr. Niko Schenk
n.schenk@em.uni-frankfurt.de

Applied Computational Linguistics Lab
Computer Science Department

Department of English- and American Studies
Goethe University Frankfurt, Germany

October 16, 2019

"Correct" vs. "Incorrect" Use of Language

Figure: Source: `http://www.elektrojournal.at/bilder/d166/Saturn_Claim.jpg`

Soo! muss Technik

Regarding the slogan four things are striking:

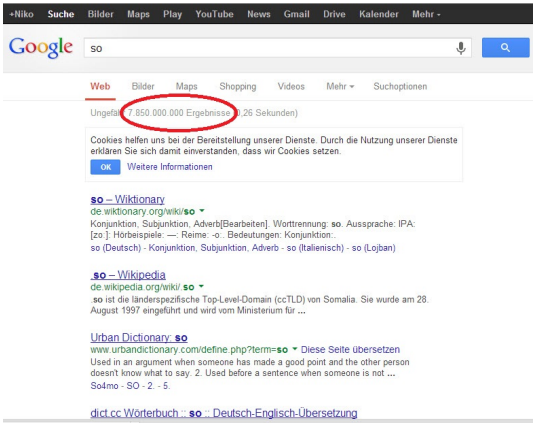
- Missing main verb "sein" (verbal ellipsis?/"ungrammatical")
- Orthographic variation "soo" vs. "so" (spelling "error")
- Misleading punctuation "soo! ...", all letters capitalized...

Question: How would you **objectively**/**formally** test that, "soo"—compared to "so"—is not "regular" language?

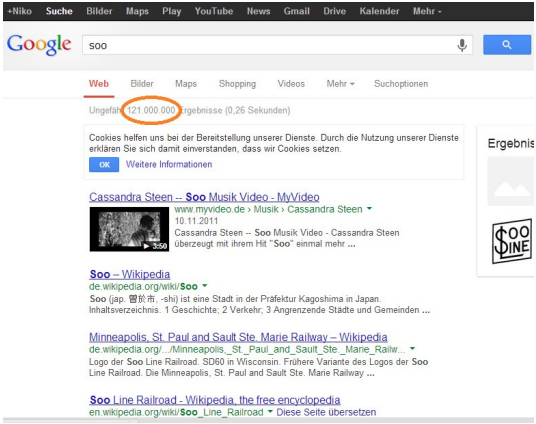
Dictionary...

Better: → `www.google.com`!

Comparing Frequencies of Contrastive Elements

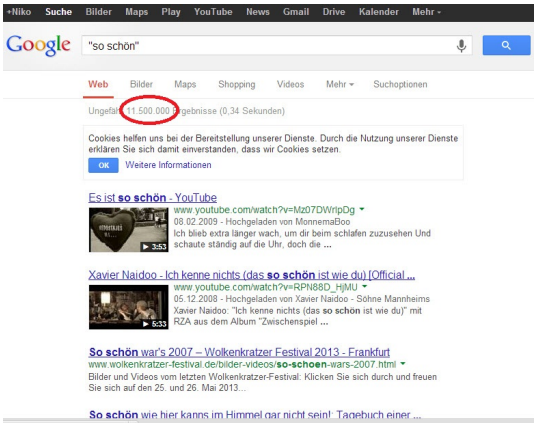


(a) Google results for "**so**"

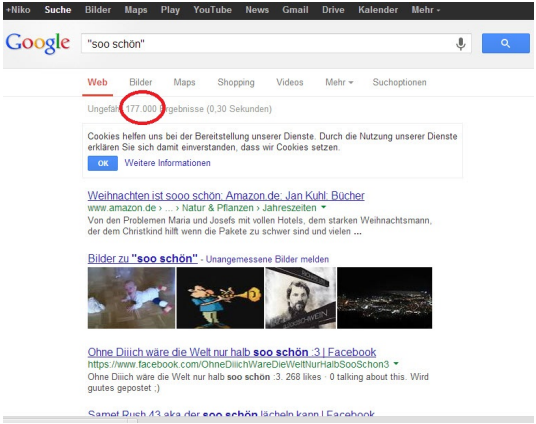


(b) Google results for "**soo**"

Refining Comparisons of Contrastive Elements



(a) Google results for "**so schön**"



(b) Google results for "**soo schön**"

Congratulations!

We've successfully performed our first **corpus linguistic search**.

How is that different from using a dictionary?
Answer:

1. consult **tons of real data** instead single of contemporary rule.
2. use **tendencies** instead of absolute true/false answer (The dictionary claims that "soo" is false or—even worse–that the phrase does not exist).

Language Change-An Example

Figure: "Thrived" vs. ...

Language Change-An Example cont'd

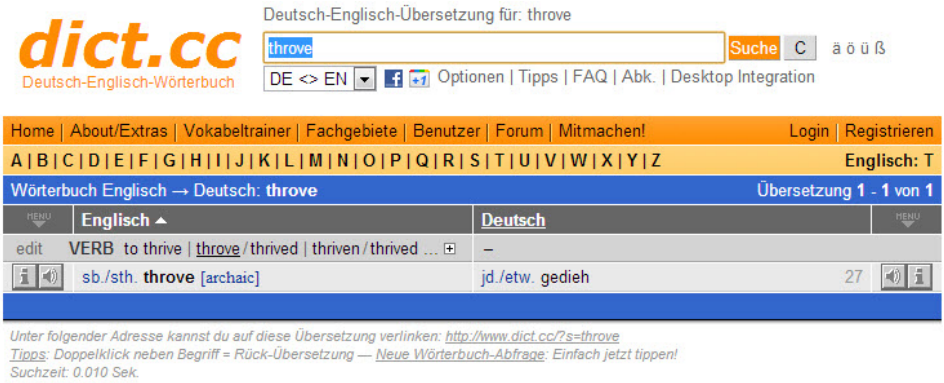


Figure: ... "throve".

Language Change-An Example cont'd

Question: How would you **objectively**/**formally** show that "throve" is obsolete/no longer in use?

You could ask a native speaker...
(Much) better: $\rightarrow$ Google books Ngram Viewer

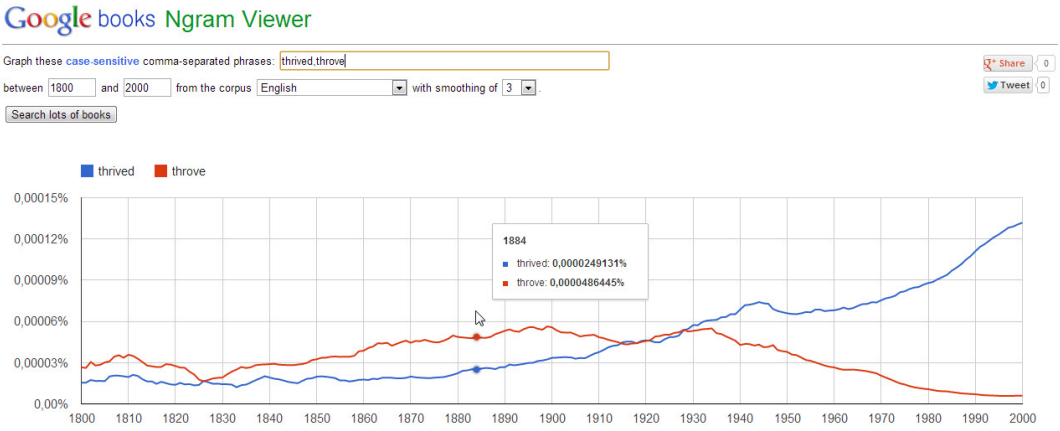


Figure: Distributions of "thrived" vs. "throve" in the Google books corpus.

Possible explanation: Low-frequency words change to fit the main paradigm.

Again, how is that different from asking a native speaker?
Answer: consult real data instead of intuition of one individual speaker.

What is Corpus Linguistics? (I)

Starting with a linguistic **phenomenon** (see previous examples) and a **hypothesis**, you use

- **large textual resources** (a corpus!) and **software** to **objectively test** (falsify/verify) the hypothesis.
- hypothesis testing is usually based on **frequencies** obtained through **search**.

Extracting Useful Information

Contrary to the previous examples, you don't have to start with a concrete hypothesis.

You could just "do something" with the corpus itself:

- e.g., compute various statistics and inspect the output.
- Usually you **count** words, phrases, etc. This is done automatically with the help of computer programs.

→ **As a result**, you can come up with a hypothesis.

Motivation

Co-occurrence probabilities between words...

Motivation


blau
00:28
82 %
@goo...
Details
Das
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das
Senden
ist
war
wird
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist
Senden
ein
eine
nicht
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter
Senden
Tag
Zustand
Zeitvertreib
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter Tag
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter Tag
Senden
für
zum
um
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

How to Automatically Find Collocations

Example: Automatically collect words which co-occur more frequently than what would be expected:

Bigram	MI-value
ramba zamba	14.39831
warminski leitheußer	14.21598
rosamunde pilcher	14.21598
nepper schlepper	14.21598
cum laude	14.21598
ante portas	14.21598
ancien régime	14.21598
aldous huxley	14.21598
solem occidisse	14.06183
popolo viola	14.06183
io muoio	14.06183
idi amin	14.06183
fata morgana	14.06183
edwin dutler	14.03366
osmoderma eremita	13.92830
hong kong	13.92830
henrich tiessen	13.92830
goenneram mappumque	13.92830
faux pax	13.92830
ping pong	13.81052

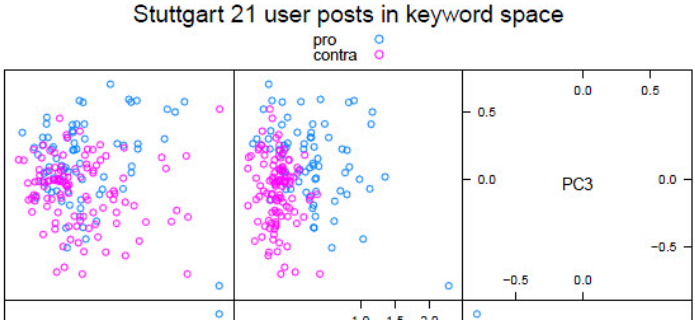
Trigram	MI-value
dierum solem occidisse	10.43753
nondum omnium dierum	10.43753
fehlgeplant und falschgemünzt	10.43753
gedeih und verderb	10.43753
clap clap clap	10.43753
ach und krach	10.43753
divide et impera	10.43753
censeo goenneram mappumque	10.28338
geteert und gefedert	10.28338
jade weser port	10.25521
rechtschreib und grammatikfehler	10.25133
ku klux klan	10.15405
cowboy und indianer	10.14984
gift und galle	10.11780
il popolo viola	10.11780
holla die waldfee	10.10105
hegen und pflegen	10.10105
pacta sunt servanda	10.03206
erstunken und erlogen	10.03206
populi vox rindvieh	10.03206

Generated hypothesis: → These words are idiomatic expressions, proper names...

How to Automatically Detect Similar Facebook Users

Based on the data, you could just count the words which are used by different people and compare the numbers.

How to Automatically Detect Similar Facebook Users



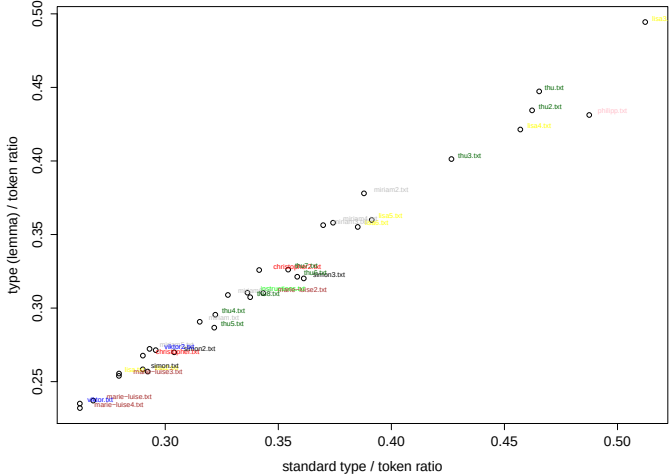
Generated hypothesis: → Groups of people with the same opinion share a similar vocabulary.

How to Automatically Identify the Author of a Text

The same technique can be applied to find the author of a document.

- For each document (written by an author), count the number of distinct words which he/she uses.

How to Automatically Identify the Author of a Text



Hypothesis Generation

Generated hypothesis:
→ Students appearing closer together in the visualization are similar in language use.

What is Corpus Linguistics? (II)

Starting with the corpus and **no** specific hypothesis, you use

- **large textual resources** and **statistics** to **detect** contrastive (interesting) patterns, i.e. you **generate a hypothesis**.

Summary

Corpus Linguistics can be divided into two parts

1. hypothesis testing
2. hypothesis generation

You usually use

- **large** textual (linguistic) resources which are **electronically available**.
- **software** to analyze (search) the data.
  - **Frequencies** are essential.

Homework Assignment

Corpus Linguistics—An Example

- Google offers an exploratory search functionality as a corpus linguistic application.
  - Cf. *Google Books Corpus*[1] / *Google Ngram Viewer*[2]

[1] `http://googlebooks.byu.edu/x.asp` – AE: 155 billion words
[2] Cf. `http://books.google.com/ngrams/`