

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



1 Hypothesis Testing

2 Hypothesis Generation

1 Hypothesis Testing

2 Hypothesis Generation

“Correct” vs. “Incorrect” Use of Language



Figure: Source: http://www.elektrojourn.al.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

Google search results for "so". The search bar contains "so". The results show approximately 7,850,000,000 results in 0.26 seconds. The first result is from Wiktionary, followed by Wikipedia and Urban Dictionary. A red circle highlights the result count.

Ungefähr **7.850.000.000** Ergebnisse (0,26 Sekunden)

Cookies helfen uns bei der Bereitstellung unserer Dienste. Durch die Nutzung unserer Dienste erklären Sie sich damit einverstanden, dass wir Cookies setzen.
 Weitere Informationen

[so – Wiktionary](#)
 de.wiktionary.org/wiki/so
 Konjunktion, Subjunktion, Adverb[Bearbeiten]. Worttrennung: so. Aussprache: IPA: [zo]; Hörbeispiele: —; Reime: -o. Bedeutungen: Konjunktion:.. so (Deutsch) - Konjunktion, Subjunktion, Adverb - so (Italienisch) - so (Lojban)

[so – Wikipedia](#)
 de.wikipedia.org/wiki/so
 so ist die länderspezifische Top-Level-Domain (ccTLD) von Somalia. Sie wurde am 28. August 1997 eingeführt und wird vom Ministerium für ...

[Urban Dictionary: so](#)
 www.urbandictionary.com/define.php?term=so Diese Seite übersetzen
 Used in an argument when someone has made a good point and the other person doesn't know what to say. 2. Used before a sentence when someone is not ...
 So4mo - SO - 2. - 5.

[dict.cc Wörterbuch :: so :: Deutsch-Englisch-Übersetzung](#)

(a) Google results for "so"

Google search results for "soo". The search bar contains "soo". The results show approximately 121,000,000 results in 0.26 seconds. The first result is from MyVideo, followed by Wikipedia and Wikipedia. An orange circle highlights the result count.

Ungefähr **121.000.000** Ergebnisse (0,26 Sekunden)

Cookies helfen uns bei der Bereitstellung unserer Dienste. Durch die Nutzung unserer Dienste erklären Sie sich damit einverstanden, dass wir Cookies setzen.
 Weitere Informationen

[Cassandra Steen -- Soo Musik Video - MyVideo](#)
 www.myvideo.de › Musik › Cassandra Steen
 10.11.2011
 Cassandra Steen -- Soo Musik Video - Cassandra Steen überzeugt mit ihrem Hit "Soo" einmal mehr ...

[Soo – Wikipedia](#)
 de.wikipedia.org/wiki/Soo
 Soo (jap. 曹於市, -shi) ist eine Stadt in der Präfektur Kagoshima in Japan.
 Inhaltsverzeichnis. 1 Geschichte; 2 Verkehr; 3 Angrenzende Städte und Gemeinden ...

[Minneapolis, St. Paul and Sault Ste. Marie Railway – Wikipedia](#)
 de.wikipedia.org/.../Minneapolis,_St._Paul_and_Sault_Ste._Marie_Railw...
 Logo der Soo Line Railroad. SD60 in Wisconsin. Frühere Variante des Logos der Soo Line Railroad. Die Minneapolis, St. Paul and Sault Ste. Marie Railway ...

[Soo Line Railroad - Wikipedia, the free encyclopedia](#)
 en.wikipedia.org/wiki/Soo_Line_Railroad Diese Seite übersetzen

(b) Google results for "soo"

Refining Comparisons of Contrastive Elements

Google search results for "so schön". The search bar contains "so schön" and the search button is highlighted. The results show approximately 11,500,000 results. The first result is a YouTube video titled "Es ist so schön - YouTube" with a thumbnail of a heart and the text "es ist so schön". The second result is another YouTube video titled "Xavier Naidoo - Ich kenne nichts (das so schön ist wie du) [Official...]" with a thumbnail of a group of people. The third result is a website titled "So schön war's 2007 – Wolkenkratzer Festival 2013 - Frankfurt" with a thumbnail of a festival scene. The fourth result is a text snippet titled "So schön wie hier kanns im Himmel gar nicht sein! Tagebuch einer ...".

(a) Google results for "so schön"

Google search results for "soo schön". The search bar contains "soo schön" and the search button is highlighted. The results show approximately 177,000 results. The first result is a website titled "Weihnachten ist soo schön: Amazon.de: Jan Kuhl: Bücher" with a thumbnail of a book cover. The second result is a text snippet titled "Bilder zu 'soo schön' - Unangemessene Bilder melden" with a row of four small image thumbnails. The third result is a Facebook post titled "Ohne Diiich wäre die Welt nur halb soo schön :3 | Facebook" with a thumbnail of a person. The fourth result is another Facebook post titled "Samet Rush 43 aka der soo schön lächeln kann | Facebook".

(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

dict.cc
Deutsch-Englisch-Wörterbuch

Deutsch-Englisch-Übersetzung für: thrived

thrived

Suche

C

ä ö ü ß

DE <> EN ▾



Optionen | Tipps | FAQ | Abk. | Desktop Integration

Home | About/Extras | Vokabeltrainer | Fachgebiete | Benutzer | Forum | Mitmachen! Login | Registrieren

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z Englisch: T

Wörterbuch Englisch → Deutsch: thrived Übersetzung 1 - 3 von 3

MENU	Englisch ▲	Deutsch	MENU
edit	VERB to thrive throve / <u>thrived</u> thriven / <u>thrived</u> ... ⊕	–	
	thrived {past-p}	gediehen 31	
	sb./sth. thrived	jd./etw. gedieh	
	sb./sth. has / had thrived	jd./etw. ist / war gediehen	

Unter folgender Adresse kannst du auf diese Übersetzung verlinken: <http://www.dict.cc/?s=thrived>

Tipps: Doppelklick neben Begriff = Rück-Übersetzung — [Neue Wörterbuch-Abfrage](#): Einfach jetzt tippen!

Suchzeit: 0.019 Sek.

Figure: “Thrived” vs. ...

Language Change-An Example cont'd

dict.cc
Deutsch-Englisch-Wörterbuch

Deutsch-Englisch-Übersetzung für: throve

throve ä ö ü ß

DE <> EN Optionen | Tipps | FAQ | Abk. | Desktop Integration

Home | About/Extras | Vokabeltrainer | Fachgebiete | Benutzer | Forum | Mitmachen! Login | Registrieren

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z Englisch: T

Wörterbuch Englisch → Deutsch: throve Übersetzung 1 - 1 von 1

MENU	Englisch ▲	Deutsch	MENU
edit	VERB to thrive <u>throve</u> / thrived thriven / thrived ... ⊕	–	
<input type="button" value="i"/> <input type="button" value="a"/>	sb./sth. throve [archaic]	jd./etw. gedieh	27 <input type="button" value="a"/> <input type="button" value="i"/>

Unter folgender Adresse kannst du auf diese Übersetzung verlinken: <http://www.dict.cc?s=throve>

Tipps: Doppelklick neben Begriff = Rück-Übersetzung — Neue Wörterbuch-Abfrage: Einfach jetzt tippen!

Suchzeit: 0.010 Sek.

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: → Google books Ngram Viewer

Google books Ngram Viewer

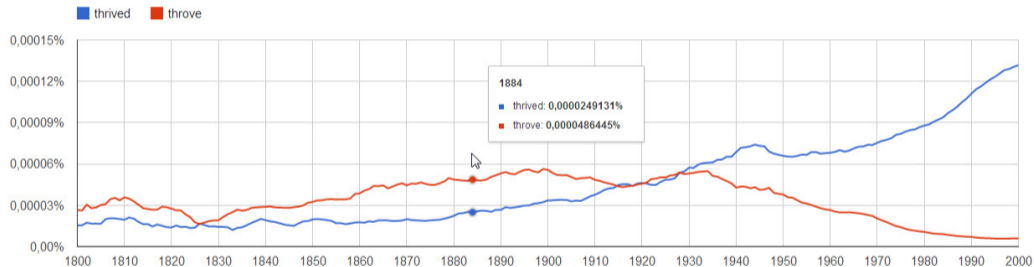
Graph these **case-sensitive** comma-separated phrases: between and from the corpus with smoothing of [Share](#) 0[Tweet](#) 0

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**.

1 Hypothesis Testing

2 Hypothesis Generation

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



Motivation



Motivation



Motivation



Motivation



Motivation



Motivation



Motivation



How to Automatically Find Collocations

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

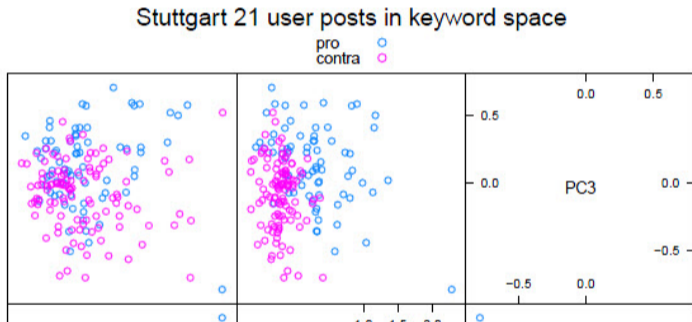
Bigram	MI-value	Trigram	MI-value
ramba zamba	14.39831	dierum solem occidis	10.43753
warminski leitheußer	14.21598	nondum omnium dierum	10.43753
rosamunde pilcher	14.21598	fehlgeplant und falschgemünzt	10.43753
nepper schlepper	14.21598	gedeih und verderb	10.43753
cum laude	14.21598	clap clap clap	10.43753
ante portas	14.21598	ach und krach	10.43753
ancien régime	14.21598	divide et impera	10.43753
aldous huxley	14.21598	censeo goenneram mappumque	10.28338
solem occidis	14.06183	geteert und gefedert	10.28338
popolo viola	14.06183	jade weser port	10.25521
io muoio	14.06183	rechtschreib und grammatikfehler	10.25133
idi amin	14.06183	ku klux klan	10.15405
fata morgana	14.06183	cowboy und indianer	10.14984
edwin dutler	14.03366	gift und galle	10.11780
osmoderma eremita	13.92830	il popolo viola	10.11780
hong kong	13.92830	holla die waldfee	10.10105
henrich tiessen	13.92830	hegen und pflegen	10.10105
goenneram mappumque	13.92830	pacta sunt servanda	10.03206
faux pax	13.92830	erstunken und erlogen	10.03206
ping pong	13.81052	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



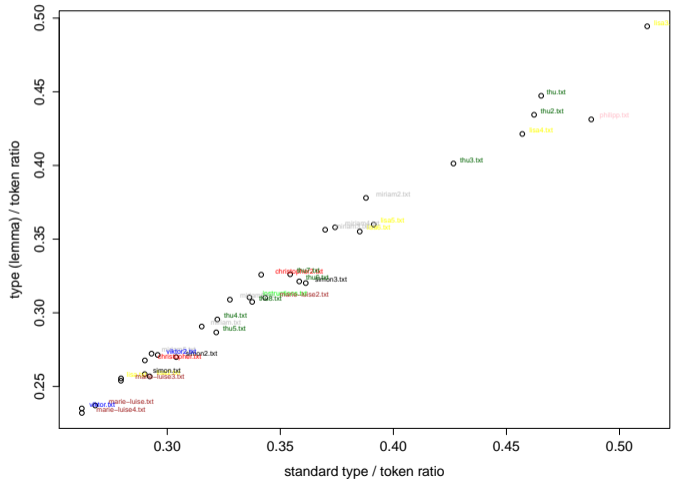
Generated hypothesis: → Groups of people with the same opinion share a similar vocabulary.

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

²Cf. <http://books.google.com/ngrams/>