Research projects: Studying lexis empirically

Seminar 'Lexicology'

Quirin Würschinger, LMU Munich

June 25, 2025

Outline

- Recap: studying lexical semantics
- Term paper requirements and consulting
- Developing research projects: topics, questions, and methods
- Research areas and empirical examples
- Data sources and analysis techniques
- Academic resources and tools
- Workshop: developing your project proposal

Recap: studying lexical semantics

In pairs (3-4 mins):

1. Think back to our semantics session. What aspect of lexicology has sparked your curiosity most?
2. What kind of research question would you like to explore?
3. What data sources (dictionaries, corpora) could help answer your question?

Connection to previous work: Remember studying differences between clippings (*admin*, *exam*, *phone*) and their source forms using collocation analysis?

Term Papers

Requirements and format

- **3 ECTS**: short paper (≈ 3–5 pages)
- **6 ECTS**: long paper (≈ 10–12 pages)
- **9 ECTS**: project report (≈ 3 pages)

Deadlines

- Short papers and project reports: **29 August**
- Long papers: **8 September**

Term paper consulting

Requirements

- Register via **email**
- Send preliminary information one day before meeting:
  - → research questions and hypotheses
  - → theoretical background, data and method
  - → table of contents and bibliography

Available dates

- 28 July (10:00, 16:00)
- 8 August (10:00, 13:00)

Zoom details

- https://lmu-munich.zoom.us/j/5385530182?pwd=SE5iZDJGQlZ1V3dpN2Q4NW45WjF5Zz09
- Meeting ID: 538 553 0182
- Passcode: 531379

Developing Research Projects

Goals for term papers

- Research question with a **linguistic focus**
- **Empirical study** using real data
- Use **corpus and/or dictionary data**
- Contribute new insights to lexicology

Your term paper should demonstrate:

- Understanding of theoretical concepts
- Ability to work with linguistic data
- Critical analysis and interpretation skills

What makes a good topic?

A good research topic:

- is **not too general** but also **not too specific**
- has further **relevance for linguistics**
- includes **new aspects** based on state of the art
- is **interesting and doable**
- is based on **previous knowledge** and/or observations
- allows for a number of **research questions**

From topic to title

Your title is the business card of your paper:

- Must be **informative and explicit**
- Must have a **reasonable link** to the content
- Must not **raise expectations** that are not met
- Often good to use a **subtitle**

Finding research questions

Research questions can come from:

1. **Previous literature**
    - suggestions for future research
    - replicating someone else's work
    - identifying gaps in knowledge
2. **Observation**
    - noticing patterns in language use
    - personal linguistic experiences
3. **Empirical findings**
    - discoveries within your own data analysis

Research questions: Key criteria

Ask yourself:

- Is it **broad enough** to be interesting?
- Is it **narrow enough** to be doable?
- Does it have a **strong enough linguistic focus**?

- What do I **expect** the outcome to be? (**hypotheses**)
- Why do I expect these results?
- How does my question relate to **previous work**?

Academic paper structure

- **Introduction**: Contextualise and motivate
- **Theoretical Background**: Review relevant literature
- **Data**: Describe your dataset
- **Method**: Explain your approach
- **Results**: Present findings clearly
- **Discussion**: Interpret and discuss
- **Conclusion**: Summarise and reflect

Research Areas and Examples

Lexical variation

How do words vary across:

- **Text types**: e.g. academic vs. social media
- **Regional varieties**: e.g. British vs. American English
- **Time periods**: e.g. historical change
- **Word-formation processes**: e.g. blending differences between registers

Example: Frequency of *autumn* vs *fall* across countries in NOW Corpus

Regional variation analysis showing British vs. American English preferences.

NOW Corpus (News on the Web)

SEARCH | CHART | CONTEXT | OVERVIEW

Frequency by country (Return to frequency by year)

SECTION	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
United States	38915	7,280.8	5.34	
Canada	10402	2,236.2	4.65	
Great Britain	78895	2,578.9	30.59	
Ireland	40637	1,257.7	32.31	
Australia	13860	1,389.1	9.98	
New Zealand	11186	680.8	16.43	
India	5923	2,022.8	2.93	
Sri Lanka	205	143.0	1.43	
Pakistan	1624	409.3	3.97	
Bangladesh	509	100.4	5.07	
Malaysia	1695	398.5	4.25	
Singapore	5889	643.6	9.15	

Semantic variation: *problem* across text types

Corpus of Contemporary American English

SEARCH | FREQUENCY | CONTEXT | OVERVIEW

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION)

SEE # TEXTS [HELP...]

SEC 1 (TV/MOVIES): 128,074,534 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	PROBLEM	990	217	7.7	1.8	4.3
2	MAN	365	12	2.8	0.1	28.4
3	KIND	237	49	1.9	0.4	4.5
4	SIR	226	0	1.8	0.0	176.5
5	DRINKING	202	55	1.6	0.5	3.4
6	PART	198	349	1.5	2.9	0.5
7	THANKS	195	1	1.5	0.0	182.4
8	HELL	171	10	1.3	0.1	16.0
9	PEOPLE	166	90	1.3	0.8	1.7
10	DRUG	150	186	1.2	1.6	0.8
11	BIT	138	5	1.1	0.0	25.8
12	TIME	132	116	1.0	1.0	1.1

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SOLVING	2255	25	18.8	0.2	96.4
2	BEHAVIOR	1208	5	10.1	0.0	258.3
3	BEHAVIORS	695	0	5.8	0.0	580.2
4	SOLUTION	610	125	5.1	1.0	5.2
5	HEALTH	374	12	3.1	0.1	33.3
6	PART	349	198	2.9	1.5	1.9
7	STUDENTS	320	7	2.7	0.1	48.9
8	PROBLEM	217	990	1.8	7.7	0.2
9	AREAS	213	16	1.8	0.1	14.2
10	APPROACH	202	6	1.7	0.0	36.0
11	DRUG	186	150	1.6	1.2	1.3
12	SKILLS	179	3	1.5	0.0	63.8

Lexical change

- **Frequency change**: How word usage changes over time
- **Meaning change**: How word meanings evolve

Example: Collocates of *gay* in COHA showing semantic shift

Corpus of Historical American English

SEARCH | FREQUENCY | CONTEXT | OVERVIEW

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION)

SEE # TEXTS [HELP...]

SEC 1 (1900): 21,977,250 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	BRIGHT	12	0	0.5	0.0	54.6
2	GRAVE	9	0	0.4	0.0	41.0
3	GAYEST	6	0	0.3	0.0	27.3
4	GALLANT	6	0	0.3	0.0	27.3
5	FRIVOLOUS	5	0	0.2	0.0	22.8
6	FRENCH	5	0	0.2	0.0	22.8
7	CHEERFUL	5	0	0.2	0.0	22.8
8	BRILLIANT	5	0	0.2	0.0	22.8
9	REAL	5	0	0.2	0.0	22.8
10	JOYOUS	5	0	0.2	0.0	22.8
11	HAPPY	8	1	0.4	0.0	12.7
12	FULL	7	1	0.3	0.0	11.1
13	LITTLE	33	5	1.5	0.1	10.5
14	SWEET	6	1	0.3	0.0	9.5
15	GLAD	10	3	0.5	0.1	5.3
16	YOUNG	18	9	0.8	0.3	3.2

SEC 2 (2000): 34,821,812 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	LESBIAN	65	0	1.9	0.0	186.7
2	BISEXUAL	15	0	0.4	0.0	43.1
3	NATIONAL	9	0	0.3	0.0	25.8
4	MALE	8	0	0.2	0.0	23.0
5	LEGAL	7	0	0.2	0.0	20.1
6	PUBLIC	7	0	0.2	0.0	20.1
7	ONLY	6	0	0.2	0.0	17.2
8	CIVIL	6	0	0.2	0.0	17.2
9	CONSTITUTIONAL	5	0	0.1	0.0	14.4
10	AMERICAN	5	0	0.1	0.0	14.4
11	MAJOR	5	0	0.1	0.0	14.4
12	MARRIED	5	0	0.1	0.0	14.4
13	RECENT	5	0	0.1	0.0	14.4
14	STRAIGHT	19	1	0.5	0.0	12.0
15	NEW	10	1	0.3	0.0	6.3
16	GAY	33	6	0.9	0.3	3.5

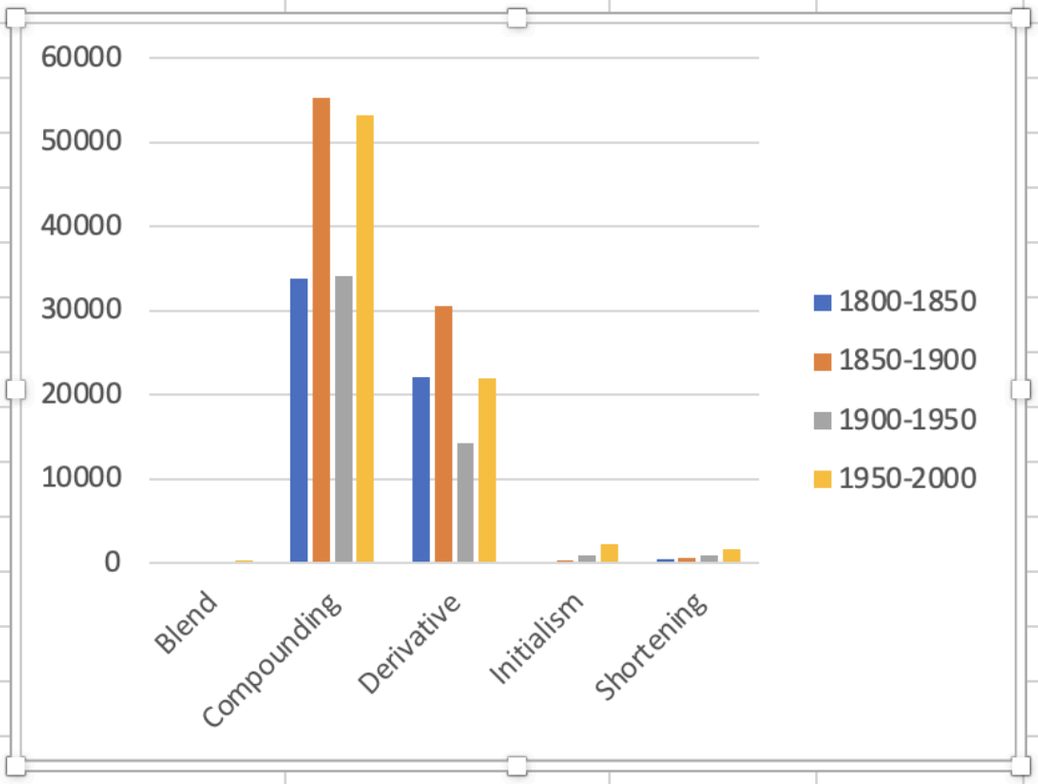
Word-formation patterns

- **Distribution** of word-formation processes across domains
- **Productivity** of morphological processes over time
- **Semantic constraints** on word formation

Distribution of shortenings across semantic domains:

Productivity of word-formation processes over time:

Sum of CountAbs	Column Labels				
Row Labels	1800-1850	1850-1900	1900-1950	1950-2000	Grand Total
Blend	34	72	153	380	639
Compounding	33843	55287	34128	53155	176413
Derivative	22069	30542	14248	21924	88783
Initialism	92	296	894	2299	3581
Shortening	431	677	881	1643	3632
Grand Total	**56469**	**86874**	**50304**	**79401**	**273048**



Data Sources and Methods

Types of data

Corpus data

- Large collections of authentic language use
- Diachronic corpora (COHA) vs. synchronic (BNC)
- Specialised corpora (academic, social media)

Dictionary data

- Historical dictionaries (OED)
- Contemporary dictionaries
- Learner dictionaries

Key corpus resources

Selected examples:

Synchronic corpora

- **BNC 1994**: British National Corpus (100M words)
- **BNC 2014**: British National Corpus (100M words)

Diachronic corpora

- **EEBO**: Early English Books Online
- **Google Books Ngram**: Historical frequency data
- **COHA**: Corpus of Historical American English (1810s-2000s)
- **COCA**: Corpus of Contemporary American English
- **NOW**: News on the Web corpus (real-time)
- English Trends: huge monitor corpus available on Sketch Engine

Corpus analysis examples

Text type variation: *brother* vs *bro* in COCA

Distribution across genres shows clear patterns:

- *brother*: more formal registers (academic, news)
- *bro*: informal contexts (spoken, fiction)

COCA analysis for *brother*:

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	118547	7794	10367	35327	12209	26215	10507	12271	3857	15688	16725	16640	17296	17216	16821
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	119.38	60.60	83.43	275.83	96.79	221.56	83.33	100.80	32.20	129.54	133.58	133.52	140.56	139.57	137.03
SEE ALL SUB-SECTIONS AT ONCE															

COCA analysis for *bro*:

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	8432	507	371	6468	229	475	207	105	70	467	579	756	1312	1988	2452
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	8.49	3.94	2.99	50.50	1.82	4.01	1.64	0.86	0.58	3.86	4.62	6.07	10.66	16.12	19.98
SEE ALL SUB-SECTIONS AT ONCE															

Academic Resources and Tools

Finding references

Academic databases:

- LLBA: Linguistics and Language Behavior Abstracts
- MLA: International Bibliography
- JSTOR: Multidisciplinary digital library
- John Benjamins e-Platform: Linguistic publications
- ScienceDirect: Elsevier journals
- Cambridge Core: Cambridge University Press

Web-based tools:

- Google Scholar: comprehensive academic search engine
- Semantic Scholar: AI-powered research tool with citation analysis
- Connected Papers: visual maps of research connections
- OpenAlex: open source academic database
- Elicit.org: AI research assistant for literature reviews

Strategies:

- **Schneeballprinzip**: Find one good reference, follow its citations
- Start with handbooks for quality overviews
- Use research network platforms (ResearchGate, Academia.edu)

Citation management

Zotero (recommended):

- Free, open-source reference manager
- Browser plugin for easy capturing
- Automatic formatting in e.g. MS Word, Google Docs
- Collaborative features for group projects

Alternatives: Mendeley, EndNote, BibTeX

Best practices:

- Start collecting references early
- Always check for citation accuracy
- Use consistent citation styles

Citation styles

For this course, use **author-date format**:

- In-text: “Corpus analysis reveals…” (Hilpert et al. 2023: 25)
- Bibliography: Consistent formatting following one style guide

Recommended guides:

- Anglistik LMU Stilblatt
- Chicago Author-Date Style
- Unified Style Sheet for Linguistics

Writing tools

Language support:

- dict.cc: dictionary search
- Ozdic.com: collocations
- Netspeak: usage patterns

AI-assisted tools:

- DeepL Write
- Grammarly
- LanguageTool: free alternative to Grammarly

Use responsibly - these are aids, not replacements for your thinking!

Reference management

Recommended: Use Zotero for:

- Collecting and organising sources
- Automatic citation formatting
- Collaboration and sharing
- Integration with word processors

Alternative: Maintain one consistent bibliography file manually

Workshop: Your Research Project

Individual task (10 mins)

Choose a research area and develop your project:

1. Select a **research area** (lexical variation, change, or word-formation)
2. Formulate a **specific research question**
3. Consider what **data** you would need
4. Think about potential **methods** of analysis

Use the worksheet to structure your thoughts.

Project worksheet

- **Topic**: _____
- **Research Question**: _____
- **Hypotheses**: What do you expect to find, and why?
- **Data**: What corpus or dictionary data would you use?
- **Method**: How would you analyse the data?
- **Challenges**: What difficulties might you encounter?

Sharing and feedback (5 mins)

In small groups:

1. Present your research question
2. Get feedback: Is it clear? Interesting? Doable?
3. Discuss data sources and methods
4. Suggest improvements

Rotate so everyone presents and receives feedback.

Getting help

Resources available:

- **Term paper consulting**: Have a meeting with me to discuss your project
- **Course materials**: Slides and readings on corpus methods
- **Library workshops**: Research skills and database access
- **Writing centre**: Academic writing support

Remember: Start early, ask questions, and use the resources available!

Summary

- **Good research projects** combine theoretical insight with empirical analysis.
- **Research questions** should be focused, linguistically relevant, and answerable with available data.
- **Corpus and dictionary data** offer rich opportunities for lexicological research.
- **Academic writing** follows established conventions - learn and apply them consistently.
- **Planning and preparation** are key to successful research projects.