12. Constructions

LoLA II: Lexical Semantics

Alexander Rauhut, M.A.

Summer Semester 2020

Contents

Welcome

Welcome to *Levels of Linguistic Analysis II: Lexical Semantics*! On this page, I will condense all presentation materials, summaries of our interactive sessions, and also the weekly homework assignments.

To get started: you can find the syllabus below with everything important accessible via links. Each homework assignment is to be prepared for the following week (#1 for week 2 ...). I also discuss the design of the course under workflow, and there is a tutorial for you to get everything ready for the course in How to set up.

You can download this whole document in .pdf or .epub[1] formats. You can also download individual chapters or download this page as html (ctrl+s) and view it in your Browser offline.

[1] Experimental, not perfect but sort of works

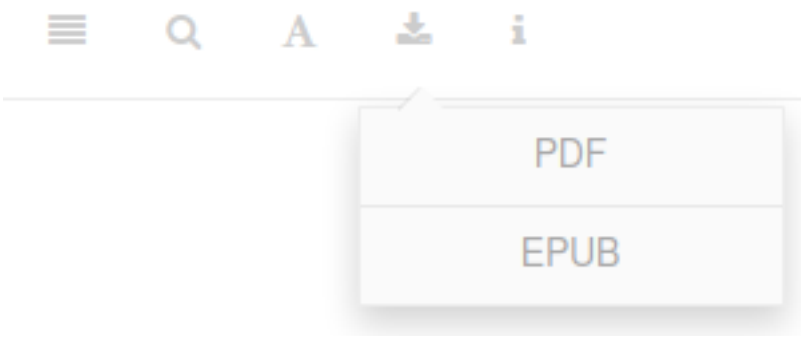


Figure 1: Just click on the download symbol in the top left corner.

Syllabus

This is not the final version!

	Date	Topic	Main Reading	Homework[2]
1	20.04.	Introduction		#1
2	27.04.	The lexicon	Geeraerts (2015)	#2
3	04.05.	Word classes	Stefanowitsch (2020) ch. 1	#3
4	11.05.	Categorization	Weisser (2016) ch. 2-3	#4
5	18.05.	Collocation	Kennedy (1991)	#5
6	25.05.	Metaphor	Deignan (2006)	#6
7	01.06.	Metonymy	Deignan (2005)	#7
8	08.06.	Synonymy	Kennedy (2003)	#8
9	15.06.	Antonymy	Justeson & Katz (1991)	#9
10	22.06.	Lexical Patterns	Altenberg & Granger (2001)	#10
11	29.06.	Lexical Bundles	Biber, Conrad & Cortes (2004)	#11
12	06.07.	Constructions	Stefanowitsch & Gries (2009)	#12
13	13.07.	Final Discussion		

Contact and Links

- Alexander Rauhut
- Email: alexander.rauhut@fu-berlin.de
- Homepage: https://alexraw.xyz
- Office Hours (Online): Monday 11-12, or whenever you catch me online.

Links

- Campus Management: Enrolment, Grades
- Primo: FU Online library
- Blackboard: Additional course materials
- StructEng Wiki: A wiki all about (corpus) liguistics written by my colleagues and me. Currently under construction.
- Oxford English Dictionary — Full access via VPN
- Prof. Stefanowitsch's Google Groups
- Tellonym: Anonymous feedback, suggestions, complaints
- Brands term paper: This repository includes all the files from my term paper live streams. Everything I am doing there will be completely reproducible and you are free to use any part of it as a template or in your project.

    - Homework
2. Live streams
    - Weekly presentations
3. Group sessions
    - Interactive sessions in smaller groups

I. This Website This website is going to be the main repository for information. I am planning for this to become a sort of blog-book-collection-of-articles-Frankenstein-monster. It will essentially replace most PDF materials you are familiar with from regular semesters, such as presentation slides. All **homework** will be published here, too. My aim is to make the experience as integrated as possible and tell the story of our class in a coherent way throughout the semesters.

II. Live streams Every week at the scheduled seminar time—Mondays 16:00-18:00—, I will live-stream my main presentation on Discord. For the most part, this will be like our regular seminar, except that we are not all in the same room. Other than that, everyone can ask questions with or without microphone, and it will be as interactive as usual (or even more so).

For now, I am not planning to upload full recordings. I might upload edited pieces from time to time to **Blackboard**. However, I will integrate anything interesting that comes up during the live session into this website. So no one is going to miss out on interesting questions or spontaneous discussions that develop during a live session.

III. Group sessions The main live sessions are going to be a bit shorter on average. This will give us time to have additional sessions in smaller groups. I have no definite plan for these yet other than having video calls in order to make everything a bit more personal. We will see how things are going after our first meeting.

How to set up

You need to set up a couple of things before we dive into the main topic. My first suggestion is to **bookmark this page**. You should also prepare to check your FU **e-mail** regularly since this is going to be the main source of news.

I. Discord You need to make yourself familiar with Discord. If you need help, check out the video below. I also discuss the setup in some more detail here in our Wiki. If you are wondering why Discord, check out the section by the same name in the appendix (or click here).

If embedded video doesn't work, click here

II. Blackboard Blackboard is mostly going to be our file storage for sensitive or copyrighted material. Over there, I will upload:

- Readings that are not available through Primo
- Material provided by or including students, e.g. student presentations / posters, recordings

Make sure you have are enrolled in this course. If you are enrolled properly via Campus Management, this should have happened automatically. Feel free to ask me for help if you are having any trouble.

III. VPN and Primo Since the library is closed, we need to take full advantage of the online resources it offers. In order to do so, every one should know how to connect to the Uni network without Eduroam. Setting up a VPN connection is part of your first homework.

1 Introduction and Organization

Main goals of week one is to get comfy with the online format and troubleshoot technical issues. We also want to refresh our memory about linguistics and get an overview about what is coming in the following weeks.

Levels of linguistic analysis I asked you:

> What did you find most interesting in the introduction?

Your Answers:

Semiotics	Phonology	Morphology	Semantics	Metaphor	Pragmatics
1	6	3	6	3	2

Rip syntax...

Linguistic Questions We are going to discuss various topics mostly from—but not restricted to—the field of Lexical Semantics, e.g.:

- What makes an **antonym**?
- How do we determine useful **collocations**, **phrases**, **synonyms**?
- How does thinking shape language? How does language shape thinking?
- What is the relationship between meaning and **grammar**?
- How can we be objective about language?

1.1 Organization

1.1.1 Course Requirements

- **Enrollment** on Campus Management (CM)
- successfully participated in an **introductory class** to linguistics

A basic grasp of linguistic concepts and a basic knowledge of linguistic terminology is required. If this is the first linguistics seminar for your, please contact me.

- **Regular participation**: Stay in touch, do homework, participate in group activity.
- **Lecture course**: regular and active participation is also required in the lecture course *Levels of Linguistic Analysis I* by Prof. Anatol Stefanowitsch.

The lecture is going to provide you with the necessary methodological knowledge and focus heavily on corpus linguistics and statistics. It is absolutely obligatory if you want to finish the course by writing a **term paper**.

If you cannot successfully join live sessions, contact me and we will find a solution. Also do so if you cannot participate in the accompanying lecture course.

1.1.2 Structure and Materials

- Essential news and official communication via **email**
- Contents on **Website**
- copyrighted material on **Blackboard**
- Interaction on **Discord**

See weekly workflow

1.2 Aims

1.2.1 Linguistic and academic skills

The introduction course had the aim to provide you with the necessary **terminology**. Like in learning a language, you need to build up your academic vocabulary before you can productively participate in any discussion. This course now is the next step. We are going to transition from reading text book chapters to actual research literature. We are going to expand the concepts and the theory behind them. And finally we are going to put it to a test by writing a linguistic study.

In the end, you will…

- Have a deeper understanding of basic linguistic concepts
- Have first experience with reading and carrying out **empirical** research
- Understand basic concepts of **cognitive science** and **usage-based** linguistics
- Understand and compile basic **statistics**

1.2.2 Skills that go beyond linguistics

Many of the skills you acquire during this class are not only useful in linguistics. Especially knowledge of empirical methodology and statistics is now more important than ever. Everyone encounters results of empirical research (good and bad) on a daily basis on the news and social media, but too few people can actually interpret the information properly. Many jobs also require at least basic knowledge in statistics.

Furthermore, there are other skills that you may benefit from indirectly, such as…

- Understanding human perception of quantities
- Understanding memory
- Understanding non-linguistic research results better
- Improve writing, reading and computer skills

1.2.3 Soft skills for Teachers

- Understand the logic behind modern teaching material
- Spot bad or obsolete material
- Understand how stubborn mistakes are learned
- Become a more aware of statistics, correlations and spurious correlations in your class room

1.3 Feedback

This format is new for most of you and some aspects of it are highly experimental. Therefore, if you encounter something that doesn't work, if you have suggestions or complaints, just let me know via any of our communication channels. Any feedback is welcome. This online semester is going to be only as good as our interaction.

You can always approach me via mail or whenever you see me online. I might be streaming some of my work from time to time if I think it might be interesting for you. Don't be afraid to join and have a chat. I am also happy to take suggestions and show you stuff that goes beyond the scope of the seminar.

1.4 Homework

In order for everyone to get used to all necessary channels, I am not providing the readings, but rather make it your first task. Now that you do not have access to the university buildings and the library everyone should learn how to connect via VPN. With a VPN connection, you have access to all online resources provided by our library.

1. a VPN connection to the university network.
   - Setup guide
2. the main readings online and download them.
   - Google Scholar (you can search for authors by typing author:*name*)
   - Primo
3. a screenshot of the directory with all your downloaded pdfs via email.

Having the possibility to connect to the university network via VPN is important even under normal circumstances. Google Scholar provides Primo links as long as you are connected to the university network (via VPN or eduroam). Every main reading can be found online.

1.4.1 Tip

I'm going to share all sorts of productivity tips for the aspiring academic at the end of every homework assignment.

Today's Tip:

Set up **shortcuts to important search engines.**

You will be doing a lot of research on google scholar, wikipedia and so on. Most browsers have some functionality to make it easier for you. Here is my setup: In my address, bar I only type 'sc keyword' or 'w keyword' and my browser searches for 'keyword' automatically on Google scholar or Wikipedia respectively (combine with Ctrl+L for hyperspeed ☺).[3] This works for most websites with a search field.

Here is where you find instructions for some popular browsers.

- For Firefox: Click here
- For Brave: Click here
- For Chrome: Click here

2 Word Classes

We are not jumping into research literature right away and have a slow semester start. The first reading was to make you comfortable again with linguistics and to recapitulate what lexical semantics is about. In the coming weeks the readings are gradually going to become longer and more technical.

Today, we are looking at the concept of words and word classes. Our aim is to provide some first evidence for common and inherently quantitative statements about language. For example, we might say that one word is "more common" than another or one word class has more members than another. Below we will start by by looking at word classes and we will try to provide evidence for a very simple hypothesis: There are closed word classes with a limited amount of members and open word classes with significantly more members.

[3] *Nachhaltigkeitsbonus*. You bypass your general search engine .

2.1 Intro

2.1.1 More organizational updates

- **Read the Instructions!**: Most problems so far boil down to not reading properly. Remember: News via mail, content and homework on course website
- **Google Groups**: Join Prof. Stefanowitsch's google groups for a forum
- **Unenroll**: If you decide to leave the course, please let me know and, if still possible, unenroll.

2.1.2 Homework discussion

Connected to the VPN, you should have found every reading except Kennedy (1991), which is available on **Blackboard**. Many important journal articles are available through **Primo** with some notable exceptions (mostly certain publishers). Many researchers also upload their articles on their websites or on platforms like **ResearchGate** Works that are much harder to get by online are monographes, and articles in collections (which are also published in books), especially older ones that are not yet digitalized. If it is just a specific quote you need, you might get lucky with a preview version on **Google Books**. Here is a collection of download links just to make sure that you get the right texts. I favored the Primo versions if there was one, but many of the texts are on Research Gate, too.

Text	Link
Geeraerts (2015)	@Primo
Stefanowitsch (2020)	–> Blackboard
Weisser (2016)	@Primo
Kennedy (1991)	–> Blackboard
Deignan (2006)	@ResearchGate
Deignan (2005)	@Primo
Kennedy (2003)	@Primo
Justeson & Katz (1991)	@Primo
Altenberg & Granger (2001)	@Primo
Biber, Conrad & Cortes (2004)	@Primo
Stefanowitsch & Gries (2009)	@ResearchGate

2.2 Parts of speech

2.2.1 Recap: Open and Closed Word Classes

The idea of open and closed word classes is the first we can quantify very easily with the help of corpus data. As opposed to a closed word class, an open word class should have a lot more members. Let's first recap what types of word classes we know.

Open word classes
- **Nouns:** *time, book, love, kind*
- **Verbs:** *find, try, look, consider*
- **Adjectives:** *green, high, nice, considerate*
- **Adverbs:** *really, nicely, well*

Closed word classes
- **Pronouns**: *I, you, she, they, mine, …*
- **Determiners**: *the, a(n), this, that, some, any, no, …*
- **Prepositions**: *to, in, at, behind, after, …*

- **Conjunctions**: *and, or, so, that, because, ...*
- ...

Closed word classes rarely accept new members. One rather recent addition to the class of **pronouns** might be considered singular *they*. Closed word classes are also mostly **invariant** in that they do not take inflection. Neither of these properties are logically necessary. You could imagine more pronouns. Some languages have a **dual** in addition to singular and plural (e.g. Classical Arabic), or a distinction between **inclusive** and **exclusive** *we* (several Polynesian languages). Yet the class of pronouns is rather fixed.

Lexical vs. Function word
- Auxiliary verbs: *be, have, (get, keep)*
- Lexical verbs: eat, sleep, repeat, ...

These first observations about word classes lead us to our core hypothesis for this week. Closed word classes have fewer members than open word classes.

2.2.2 PoS-Tags

Figuring out the word class of each word is done with **Part-of-speech taggers**. Tools like the *Tree Tagger* (Schmid 2013) can determine word classes with an accuracy of around 95% (Horsmann, Erbs & Zesch 2015). Even though this is good enough for most purposes, you have to bear in mind that automatic annotation is error prone and can cause some spurious patterns that have to be accounted for. We will encounter such cases in future sections.

PoS-tags
- annotation for word class available in most corpora
- automatized
- around 95% accuracy (Horsmann, Erbs & Zesch 2015)
- e.g. *Tree Tagger* (Schmid 2013)

2.3 Types and Tokens

2.3.1 Word boundaries

We have to make a first technical distinction at this point. We need to decide what we count as a word. In corpus linguistics, the word model most commonly encountered is the **token**. A token has a very rough and technical, yet simple definition.

Token
- character sequences in between spaces

The emphasis here lies on **character sequence**. If we use this to count occurrences we are dealing with the related concept of **type**.

Type
- class of identical tokens

Note that neither relying on spaces nor on orthographic characters is by itself ideal in most circumstances. The terms *type* and *token* are sometimes also used much more abstractly. You could understand types and tokens as a "words" disregarding spelling conventions. This requires some more work defining *word* and also working with data later.

How many words? The concept of word is actually very hard to define and its definition depends on several factors. Consider the following data:

(1) living room

living room has a coherent meaning that is highly conventionalized and also culturally specific. It is not purely **compositional**. It contrasts **paradigmatically** with words like *kitchen, attic, bathroom*, which are either clearly **monomorphemic** or at least orthographically presented as one word. Semantically, you might decide to consider it one unit rather than two. This is not necessarily true for a morphological perspective.

(2) mother-in-law

Semantically, we have a similar situation to the example above. However, we can make the observation that the plural can attach to the first component, thus *mothers-in-law*. We also find *in-laws*. The examples below demonstrate that there is some variation in where speakers feel the word boundaries are. Note that the Oxford English Dictionary (OED) (2020) recognizes *mother-in-laws* as a rare variant.[4]

(3) They wore it only because their **mothers-in-law** insisted. (BNC[5])
(4) I always thought it was **mother-in-laws** that cause the problem. (COCA[6])
(5) Angela sided with her new **in-laws**. (BNC)

Next we have fixed grammatical expressions, which are written as separate words, but mostly understood as one word:

(6) *going to*
(7) *in spite of*

going to is undoubtedly one word in spoken language (*gonna*). *in spite of* again contrasts paradigmatically with words spelt as one, such as *despite*. Especially prepositions and conjunctions in English have rather arbitrary spacing; consider for example *nevertheless*, *however*.

In summary, the definition of word strongly depends on the point of view.
You might distinguish:

- **Orthographical words** (mostly congruent with *token*)
- **Phonological words**
- **Morphological words**
- **Lemmas**

2.3.2 Word classes in numbers

Now let's turn back to our hypothesis that there are open and closed word classes. The evidence we need is **counts** for words and word classes. In an electronic corpus, the notion of orthographical word is the easiest to begin with. We basically count everything surrounded by spaces as a unit, a **token**. Below I show you the commands used to retrieve the data from our version of the British National Corpus (2007) . You don't need to worry about it just yet. In the first lessons I will provide the data and the numbers. The code might be interesting for you at a later stage, however.

```
BNC> [pos = "NN.*"]                          # get all tokens tagged as noun
BNC> count by hw > "noun_types.txt"          # count every lemma and save as .txt file
BNC> exit                                    # exit cqp and use wc -l (count lines)
$ wc -l noun_types.txt                       # repeat for other word classes, (V.*, AJ.*, AV.*, CJ.*)
```

In fact, we find that the open word classes do have considerably more **types** than closed word classes. Not a very exciting result, and not one we would necessarily need corpus linguistics for, but, nevertheless, our first empirical evidence for a linguistic concept.

[4] "mother-in-law, n. and adj." OED Online, Oxford University Press, March 2020, www.oed.com/view/Entry/122659. Accessed 28 April 2020.
[5] British National Corpus (The British National Corpus 2007)
[6] Corpus of Contemporary American English (Davies 2008)

PoS	Tokens	Types
Nouns	21255608	222445
Verbs	17870538	37003
Adjectives	7297658	125290
Adverbs	5736409	8985
Prepositions	11246423	434
Conjunctions	5659347	455
Articles	8695242	4

An observation that was not immediately apparent is that function words, though there are not too many, are very frequent individually.

- Function words have a low **type frequency**
- Function words have a high **token frequency**

In fact, the most frequent **tokens** in a corpus are function words. Below, I retrieved the 100 most frequent lemmas from the BNC.

```
$ cwb-scan-corpus BNC hw | sort -nr | head -100 > "bnc_lemma_freq.txt"
```

```
##     rank     lemma   count
## 1      1       the 6043904
## 2      2         , 5017057
## 3      3         . 4715138
## 4      4        be 4121794
## 5      5        of 3041843
## 6      6       and 2617879
## 7      7        to 2594667
## 8      8         a 2165370
## 9      9        in 1938587
## 10    10      have 1317166
## 11    11        it 1215335
## 12    12        he 1198489
## 13    13         i 1146605
## 14    14      that 1119424
## 15    15       for  879034
## 16    16      they  842806
## 17    17       you  805600
## 18    18         ‘  770022
## 19    19       not  767849
## 20    20         ’  752178
## 21    21        on  729963
## 22    22      with  658980
## 23    23       she  654447
## 24    24        as  653874
## 25    25        do  538288
## 26    26        at  521903
## 27    27        by  512381
## 28    28        we  504575
## 29    29      this  453739
## 30    30        's  447030
## 31    31       but  446125
## 32    32      from  425198
## 33    33         )  397970
```

```
## 34    34         (  391974
## 35    35         ?  387952
## 36    36        or  367091
## 37    37     which  365427
## 38    38        an  336953
## 39    39      will  336274
## 40    40     there  319390
## 41    41       say  318439
## 42    42       one  306169
## 43    43     would  278613
## 44    44       all  277131
## 45    45         -  272488
## 46    46       can  263372
## 47    47         :  257173
## 48    48        if  253331
## 49    49      what  240426
## 50    50        so  239228
## 51    51        go  229103
## 52    52        no  226862
## 53    53       get  213555
## 54    54      make  210829
## 55    55      when  209620
## 56    56      more  209561
## 57    57        up  207862
## 58    58         ;  202801
## 59    59       who  200710
## 60    60       out  197182
## 61    61     about  191813
## 62    62       see  185693
## 63    63      time  181640
## 64    64     other  181517
## 65    65      know  178347
## 66    66      take  173930
## 67    67      some  167127
## 68    68      year  161657
## 69    69     could  159880
## 70    70      into  157672
## 71    71      well  156761
## 72    72      like  155992
## 73    73      then  154587
## 74    74 [unclear]  152636
## 75    75       two  152616
## 76    76      only  148562
## 77    77     think  145729
## 78    78      come  144812
## 79    79      than  144656
## 80    80         !  141764
## 81    81         "  141750
## 82    82       now  139125
## 83    83       use  138213
## 84    84      over  130787
## 85    85      good  128684
## 86    86       may  127317
## 87    87      work  126958
```

```
## 88    88      just  126333
## 89    89      give  126225
## 90    90       new  124994
## 91    91     these  123492
## 92    92      also  123389
## 93    93    people  123259
## 94    94       any  121795
## 95    95     first  120712
## 96    96      look  120569
## 97    97      very  119437
## 98    98     after  113788
## 99    99       way  110441
## 100  100    should  109024
```

2.3.3 Considerations

There are more things to consider when counting word types. Words might only be spelt the same by coincidence, we might have words in multiple word classes, words with different senses, etc.

- Homonyms
- Polysemy
- Conversion
- Prototype theory

> (…) the distinction between V[erbs], A[djectives], and N[ouns] is one of degree, rather than kind (…)
>
> — Ross (1972)

2.4 Outlook

Next week we will have a closer look at the lexicon, especially **lemmas**. We will also refine our idea of word class and at the same time generalize our idea of a linguistic category.

- Models and categories
- A word class as a model
- A word class as a category

Also - Groups and presentation

2.5 Homework

Follow-up reading: Gries (2009)
Download Link

As a follow-up I can warmly recommend the following interview (2009) with Stefan Gries, who is easily one of the most important corpus linguists at the moment (thanks to Helen for the tip).

2.5.1 Task

Our weekly homework will mostly deal with presenting data. We are going learn how to discuss linguistic examples and visualize numbers in graphs and charts.

Linguistic examples Every linguistic discussion should present and contrast linguistic categories with the help of authentic examples. There are 3 things to consider:

1. Examples should be consecutively numbered throughout a text.
2. If you refer to a word or phrase inside your main text, it should be in italics (kursiv).
3. You should prefer authentic examples over invented examples.

Consecutive means that every example has its unique number starting at (1). If you introduce a new example later in the text, the counter continues. For a demonstration, just look at the previous section of this website. The advantage is that you can refer to examples easily by referencing its number. Another advantage is that you can use keys for cross-referencing rather than a literal number, which allows you to change the order of examples or introduce new examples before existing ones without messing up the numbering. In *Latex*, there is a package called *gb4e*, which offers an environment \ex for this purpose. It is a bit more cumbersome in something like *Word* (search for **cross reference**), but still possible. You can, of course, also do it manually, but it's worth learning.

Authentic refers to examples that have actually been produced by a (native) speaker. It is preferable to provide a reference for your example rather than to make it up yourself. Sometimes examples are invented out of convenience or to illustrate uncontroversial structures, but you are on the safer side if you provide an authentic source.

Today's task is the following:

- Provide a short definition of **homonym** and illustrate it with examples.
- Contrast it against the concept of **polysemy**.

Examples should be complete phrases or sentences and respect the conventions listed above. This assignment shouldn't exceed half a page. Be concise. Send me the assignment via email as a .pdf document. You don't need to know yet how to search for corpus examples, you can simply use dictionaries or search online for examples.

2.5.2 Tip of the day

Today's tip is from the category: **Th**ings **I w**ish **I** had **l**earned **be**fore **m**y **Ba**chelor Thesis
In short: *Tiwilbemba*.

Build your personal .pdf library

Take every .pdf you download and get from your instructors and archive it with a naming scheme you can remember easily. Especially scans from books and collections are an invaluable resource for reasons discussed above.

My suggestion: lastname_year_keyword: e.g. Deignan_2005_Metaphor.pdf

Also…

Start building your bibliography database

Get the info for a bibliography entry as soon as you read a text. In a future installment of *Tiwilbemba* I will discuss the benefits of tools like BibTex, Mendeley, Endnote …

~15s invested per text, hours saved in the long run.

---

3 The Lexicon

Last week we took a very simple idea—that there are open and closed word classes—and tried to provide evidence to support this. That created the need to properly count words, thus **quantify** a linguistic category.

In the following sections we will have a look at some key concepts in lexical semantics, and then explore some first frequency patterns that provide insight into language and cognition.

3.1 Intro

3.1.1 Organizational notes

- Lecture!
- Announcements for student presentations at the end.

3.1.2 Recap

Important Concepts

Indicator	Linguistic Concept
Tokens	word, slot (syntagmatic)
Types	word of the same form, (paradigmatic)
Parts of Speech	word class
Frequency	commonness, salience, …
Lemma	Lemma

Always remember:
Most linguistic categories can only be quantified indirectly.

What counts?
Look at the frequency list with all *lemmas* in the BNC corpus. Did you spot anything weird? We talked about representations of linguistic concepts in corpora. The best example of how orthography-centric corpora are (necessarily), is **tokenization**. Most corpora are designed so that punctuation is treated as individual tokens. Also **clitics** such as the possessive *'s* and contracted forms of auxiliary verbs such as *'ll, 've, 're* are treated as separate tokens. This decision might be contrary to the definition of word you are working with.

3.2 Lexemes and lexical fields

3.2.1 Lemma

What are all the grammatical forms of *be*, *cut*, *tree*, *nice*, *beautiful*?

(8) be, am, are, is, were, was, been, 's, 'm, 're, ?being
(9) cut, cuts, (cut, cut), ?cutting
(10) tree, trees, tree's, trees'
(11) nice, nicer, nicest
(12) beautiful

A lemma is all the **inflectional** forms of a word. This includes forms with grammatical affixes (*tree*, *trees*) and **suppletive** forms (*go*, *went*). What is not included is **derivational** suffixes like the adjectival *-ly*. Of course, this requires a clear definition of inflection and derivation. Some researchers might argue that the participial *-ing* is derivational rather than inflectional.

There is also the issue of whether the past participle of some verbs like *cut* is to be seen as separate "form" or not.

When it comes to the technical side of research, you have to be aware of the decisions taken when lemmatizing corpus data as to what counts and what doesn't. A lemma in a corpus is not equal to a lemma as a linguistic concept.

3.2.2 Distribution

How can we find out if something is a homonym if we do not know the meaning or want to keep intuition out of the picture?

Animal or sport utensil?

(13) Maybe I'm a fruitarian **bat**
(14) … with a straighter **bat** than some of the Englishmen
(15) The unfortunate starved **bat** was then returned
(16) And not simply a bat, but an autographed **bat**

(examples from The British National Corpus 2007)

```
BNC> [pos = "AJ.*"] []? [hw = "bat"]
```

In this example, the preceding adjective provides enough context to disambiguate the two meanings. If you expanded this to more co-occurrence patterns, e.g. with verbs or even different text types, two clearly distinct patterns emerge. The animal *bat* eats, like other animals, whereas the utensil *bat* strikes like other club-like devices. A Giraffe rarely strikes and a tennis racket doesn't eat. They each form distinct lexical **fields**. **Distribution** plays a defining role in the structure of our lexicon.

3.2.3 Association

A key component of human memory is association. The lexicon is organized in associative networks, **semantic fields**. What we **perceive** together frequently, we associate as belonging together. This is also referred to as spatial or temporal **contiguity**.

(17) law and …?
- order

(18) good or …?
- bad, evil

(19) the number of the …?
- ??beast

(20) spoils of …?
- ??war

The first word that comes to mind when you read the first two fragments is most likely *law and order*, and *good or bad*. For the other two examples, there is expected to be more variation. A metal fan might readily come up with *beast*, since the song of the same name is part of their cultural experience, and therefore, very frequent for them. *spoils of war* might not be a phrase that everyone is familiar with at all. *spoils* as a word is very rare; yet there is a strong association with the phrase. If it is encountered, it occurs together with *war* more often than not.

3.3 Frequency and memory

3.3.1 Common and uncommon vowels

In order to illustrate some basic frequency effects (as in count not pitch), we had a little experiment in class today with German vowel sounds.

Let's take a subset of the German monophthongs with relatively consistent phonetic spellings. We're taking orthography as an approximation for pronunciation here.

Vowel	Frequency counts (DWDS[7])	Perceived difficulty	Experimental counts
/iː ɪ/	267,353	easy	56
/aː a/	162,873	easy	37
/uː ʊ/	113,065	easy	53
(ü) /yː ʏ/	30,568	difficult	22
(ö) /øː œ/	24,562	difficult	30

Experimental task: Find as many adjectives as you can that contain the given vowel (long or short) within 3 minutes.

The expected outcome: People find most words with *i*, then *a* and *u*, and much less with *ü* and *ö*. We can see a **correlation** with the frequency with which those vowels appear in corpus data and how difficult learners find their pronunciation. The extra-ordinary performance of our *u*-group can partly be explained by the participants discovering adjectives with the very productive prefix *un-*. This in it self is an interesting association pattern.

It makes sense to hypothesize that it is easier to come up with examples if there is more to choose from. Furthermore, we can observe that front rounded vowels are rare across language (Maddieson 2013). But why are those vowels so much rarer in the first place?

Their are three obvious possibilities:

- We made a mistake
- It is coincidence
- There is something categorically different about *ü* and *ö*

Let's assume the latter is the case. What *ü* and *ö* have in common is that they are front rounded vowels. In fact, we have a pretty good idea about why they are special. In a nutshell: the frequency make-up of front rounded vowels is not as distinctive as the one's of other vowels. [a, i, u] are extremely distinct from each other so (almost) all language make a distinction between them. [i] and [e] are more similar in sound yet still much more distinct than [i] and [y]. It is much more common to see a language make a distinction between the former than the latter. The exact cross-linguistic patterns and the interesting bio-phyisical reasons are far outside the scope of this course. The important conclusion is that we found an interesting correlation with the help of corpus data that we could corroborate with other pieces of data, and that ultimately leads us to a fundamental property of language.

3.3.2 Confounding variables

We measured vowel counts with orthographic characters.
What could skew our data systematically?

- *i*, *a*, and *u* occur in diphthongs
- *i*, *a*, and *u* might represent different monophthongs (especially in loan words)

[7] DWDS Kernkorpus 21 (2000–2010); example query: `*ö* WITH $p=ADJ*`

- *ö* and *ü* are sometimes transliterated with *oe* and *ue*
- ...

There are always many factors that could skew your data in one direction or another. In this case, the observed pattern is probably amplified by the variables above. Ideally, you would control for those confounding variables, and if you can't, judge the potential implications.

3.4 Homework

Follow-up reading: Berg (2000)
Download Link (VPN)

This is a very interesting study for those of you who are interested in Phonology. It illustrates nicely how word classes are continuous categories and how to investigate this idea empirically. It is a bit of an advanced read but worth skimming through.

3.4.1 Presentations

Beginning in week 5, we will have student contributions and also more group based work. For your contribution, you can choose from the following options:

a. **Live presentation**: *2–3 people per topic*
   Present a topic related to one of the weekly readings, weeks 5 to 12, in form of a short talk. The presentation should contain a primer for **discussion**, and an attempt at **reproducing** parts of the data in the reading with our current methodological knowledge. For the exact format, you can be creative.
   - You can do a screen share with a slide presentation, or
   - Send a handout and talk on camera.
   - You can even shoot a video for us to watch beforehand if you feel like.[8]

b. **Poster presentation**: *1–2 people per topic*
   Develop your own idea for a research topic and present it in the form of an academic poster. Ideally, the topic is already close to the research question of your term paper, but that's not a must. The posters are to be submitted before the last week (13.07.) and presented during the last session in short 5-minute pitches with subsequent discussions. There is also the option of doing it asynchronously in form of a video submission. I will upload more information about how to create an academic poster when the time comes.

3.4.2 Task

1. Pick either one of the options listed above. You can also be creative and send me your suggestion.
2. If you would like to prepare a live presentation, please let me know via email, which two topics you are interested in. Provide a first and a second wish. If you already have a partner or group, let me know.
3. Watch this video.

3.4.3 Tip of the day

When you write homework, essays, term papers, or even presentations, keep writing and formatting separated. Pick a pre-made document template, and stick to it. Don't customize, don't build from scratch. Keep your formatting at a bare minimum!

[8] Provided you send it early enough so everyone can watch it before class.

In academic writing across disciplines, all the different style guides you have to deal with might be overwhelming and confusing. But in the end, it can all be boiled down to just three key elements: text, data, and reference.[9] Only the first two need to be taken care of manually during the writing process.

Text should be arranged in coherent paragraphs. Section headlines should have some specific formatting so they can be used as key for a table of contents or cross-referencing. Your type setting tool of choice (*Word* for most) has a way to deal with this; learn it! Anything else should be taken care of by your template.

When it comes to presenting **data**, here are the only three elements you should bother with manually.

- Meta-linguistic reference: words and phrases as in-text examples in *italics* (see last homework)
- Listed examples (see last homework): indented and in their own paragraph, consecutively numbered
- Tables and figures: keep it simple here, too. They need to have title, numbering and description. Don't bother applying unnecessary visual effects, or having the text flow nicely around them. If the table or figure doesn't fit, it belongs in the appendix.

In a well written text, you don't need any other visual emphasis, except maybe to highlight parts of listed examples in **bold**. Italics, underlined or colored text is otherwise unnecessary. There are also long quotes, book or journal titles, footnotes, and listings; however, it's worth considering whether you actually need them. In most cases, you are better off skipping those.

Then, there are tables of content, citations and bibliographies, cross-references lists of tables/abbreviations etc.; but here is a simple rule I learned the hard way: **Never** create these manually—**never**! There are ways to deal with citations and bibliographies automatically that allow you to apply whatever style your instructor or potential publisher requires. I will return to them in a future Tiwilbemba.

In summary, keep things simple, be aware of the elements in your text, and don't mix. Extensive formatting can be a huge time sink and should be avoided.

4 Categorization

4.1 Intro

4.1.1 Recap

Last week, we explored some important concepts related to the lexicon and some cognitive processes determining them.

Lexical structures

- Lexeme: set of inflectional forms that are related via their meaning
- Lemma: all inflectional forms of a lexeme
- Lexicon: system of lexemes
- Lexical field: class of lexemes with common meaning and co-occurrence patterns

Cognitive concepts

- Association: emergent networks in memory
- Contiguity: spatial and temporal correlation
- Salience: distinctiveness

[9] This list is specific to linguistic articles but the principle applies to most pieces of text

4.2 Models

4.2.1 Simplify, Generalize, Apply

- A model is supposed to **simplify** the complexity of reality
- Explain and approximate reality as well as possible with as few concepts as possible
- These generalizations should be useful in
  - further research
  - practical application

4.2.2 Some Linguistic Models

- **The linguistic sign** (Saussure)
- **The Lexicon**:
  - a model of how auditory and visual stimuli are organized in memory
  - adds information about range of closely related forms
  - adds information about syntagmatic and paradigmatic relationships
- Prototype Theory:

4.2.3 Short and long vowels

Sometimes a certain way of describing a phenomenon sticks around even though it is incomplete and sometimes even false. An interesting example can be found in 'long' and 'short' vowels in Germanic languages.

- long vowels: /ɑː/

This is where it makes sense to think of it in terms of models.

4.2.4 Voiced and voiceless consonants

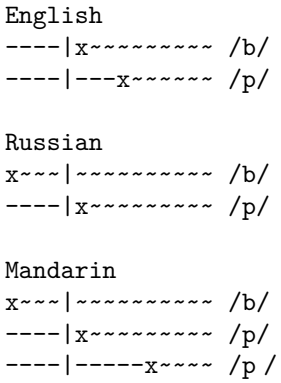
Another example is voicing of plosives in Germanic languages. Traditionally, the differences between /p/ and /b/ is described as one of voice.

- voiceless plosives: /p, t, k/
- voiced plosives: /b, d, g/

Alternative terminology: fortis/lenis, strong/weak, aspirated/non-aspirated

Dichotomies like the ones above, however, never capture the full complexity of a linguistic phenomenon.

Voice onset time



4.2.5 Expanding the model

Our **model** of voicing is becoming more and more complex. Even voice onset time is not the full story. You can also observe specific patterns in the phonological environment such as assimilative voicing, or vowel length of adjacent sounds. In a Russian accent, the phrase *it's better* might be pronounced as [ɪdz betə] as opposed to a more native [ɪts betə]. This pattern is clearly connected to the different voice onset times, but is not a logical consequence. We can enrich our model even more:

- devoicing of following continuants (progressive assimilation)
  tree [t̥i:]
- lengthening of preceding vowels
  bet [bet]—bed [beˑd]
  beet [bi:t]—bead [biːˑd]

4.2.6 Different models for different purposes

Models differ substantially in focus and the degree of detail. A simple dichotomy like voiced/voiceless consonants and long or short vowels is only a rough approximation of reality but at the same time, it could be just enough for a certain context. In fact, the very purpose of a model is to reduce complexity. Let's consider the following applications.

- Teaching orthography to native speakers
- Teaching pronunciation to language learners
- Explaining historical sound changes
- Articulatory phonetics
- Speech recognition

In native language teaching, the aim is often to teach spelling conventions. Since a native speaker doesn't normally need any instruction in how to produce speech sounds, any categorization that does not lead to confusion will do in order to distinguish vowels or consonants. The issue in foreign language teaching, on the other hand, is a completely different one. Students might need instruction on how to produce the difference between the sound categories. Here the level of detail needed even depends on the native language of the student. If you teach English to a German student, the simplest of the models above might be enough since the patterns in the languages are very similar. However, a Mandarin or Russian speaker might profit from a more complex model to explain the difference. In a linguistic context, the models are of course much more complex, but even here you'll find differences. A historical linguist focusses on different aspects than a phonetician. Then there are purposes like speech recognition where the type of model might become yet more detailed. When trying to capture speech with a computer, you need to model the sound differences with frequencies (as in pitch). This most extreme model makes it useful for its specific task but renders it useless for most of the others.

5 Collocation

Coming soon

5.1 Homework

5.1.1 Task

From the next reading (Deignan 2006), pick out some tables with frequencies and visualize them. There are several ways to visualize count data and proportions. The most common

ones are bar charts and stacked bar charts. Your task is to decide which way is best and to figure out how to do it. Most of you would probably want to do this in Excel, Libre Calc or another spreadsheet program of your choice. Consult your favorite search engine. :)

In a nutshell:

1. Pick some tables from the reading that you find interesting.
2. Decide about the best way to visualize them.
3. Learn how to do it and send me your figures in .pdf format.
4. (Expert mode: Do you think the data is best presented on a log-scale?)

5.1.2 Tip of the day

Here are two seemingly unconnected thoughts on co-occurrence patterns and exponential decay (Zipf's law).

Fact 1: A place where n-grams and co-occurrence patterns were used to make something useful is the computer keyboard. The keyboard layout (QWERTY) was carefully designed to avoid the most frequent letter combinations (bigrams and trigrams) to be on adjacent keys (oversimplified) so that old mechanical typewriters don't get jammed.

Fact 2: When you work on a project, the amount of time you use on individual aspects also follows a power law like Zipf's law. Look up the *Pareto principle*. You probably need 80% of your time to produce 20% of the work and 20% of the time to produce 80% of the work. You cannot avoid that, but you can flatten the curve by focusing on the biggest time sinks, e.g. by following my formatting tips.

Loosely related, my tip of the day is another tiwilbemba: Learn touch typing if you haven't already. Since you study language, chances are, you will spend most of your work time typing. Learn a good, fast and healthy typing technique and you can save a lot of time in the long run. Just imagine how much faster you'll write your Bacholor Thesis if you type at twice the speed mistyping half as often, which is easy to achieve for most people. I wish I had learned that many years ago. Believe it or not, it can actually be fun. If you have learned a musical instrument—this is basically what it is like, just much faster to master. If you haven't learned a musical instrument—forget about touch typing and learn an instrument. :D

6 Metaphor

This week we will have a brief look at metaphor. When we think of metaphor, we usually have literary metaphor in mind. However, as soon as you try to give a systematic definition of metaphor, you will notice how pervasive the concept is. As a matter of fact, many lexemes have common metaphorical uses and it becomes very difficult to draw the line. One major way that our lexicon is enriched with new meanings and uses of existing lexemes is by **metaphorical extension**.

Examples:

(21) *to light up*

a. *Why should he **light** up his front lamp to time?* (BNC)
b. *My eyes **light** up at the sight of her.* (BNC)

(22) *ocean*

a. *Only a little earthy bank separates me from the edge of the **ocean**.* (BNC)
b. *The smaller yurt was an **ocean** of coolness and quiet.* (BNC)

Sometimes metaphorical uses are correlate with certain lexical and grammatical contexts. *light up* for example is used metaphorically whenever in the context of *face* or *eyes*. The construction [an ocean of x] is almost always used metaphorically to express an extremely large quantity of x. x in this case is itself something resembling a liquid substance only via metaphor. Emotions are often understood as liquids. Today's reading assignment Deignan (2006) explored some metaphorical patterns relating to singular and plural nouns.

6.1 Metaphor and quantitative evidence

6.1.1 Coding

One of the main challenges concerning metaphor in corpus linguistics is that it is hard, and sometimes impossible, to extract metaphorical uses automatically. The following list highlights the major implications of this.

- Manual coding is time-consuming
- Manual coding is error prone; requires rigorous operationalization
- Frequencies of metaphorical uses are often dwarfed by non-metaphorical uses
- There is often no way to distinguish "literal" and "metaphorical"

6.1.2 Homework discussion

Before I get into a summary of the main indicators we've used so far, here is a disclaimer first. The data from Deignan (2006) was very limited. The frequencies are extremely low and all the transformations and visualizations you can apply can easily become misleading. Generally, I advice against over-analyzing low counts. The purpose of the homework was to make you experiment on a very simple data set rather than to learn something new about the data.

6.1.3 Frequency and scales

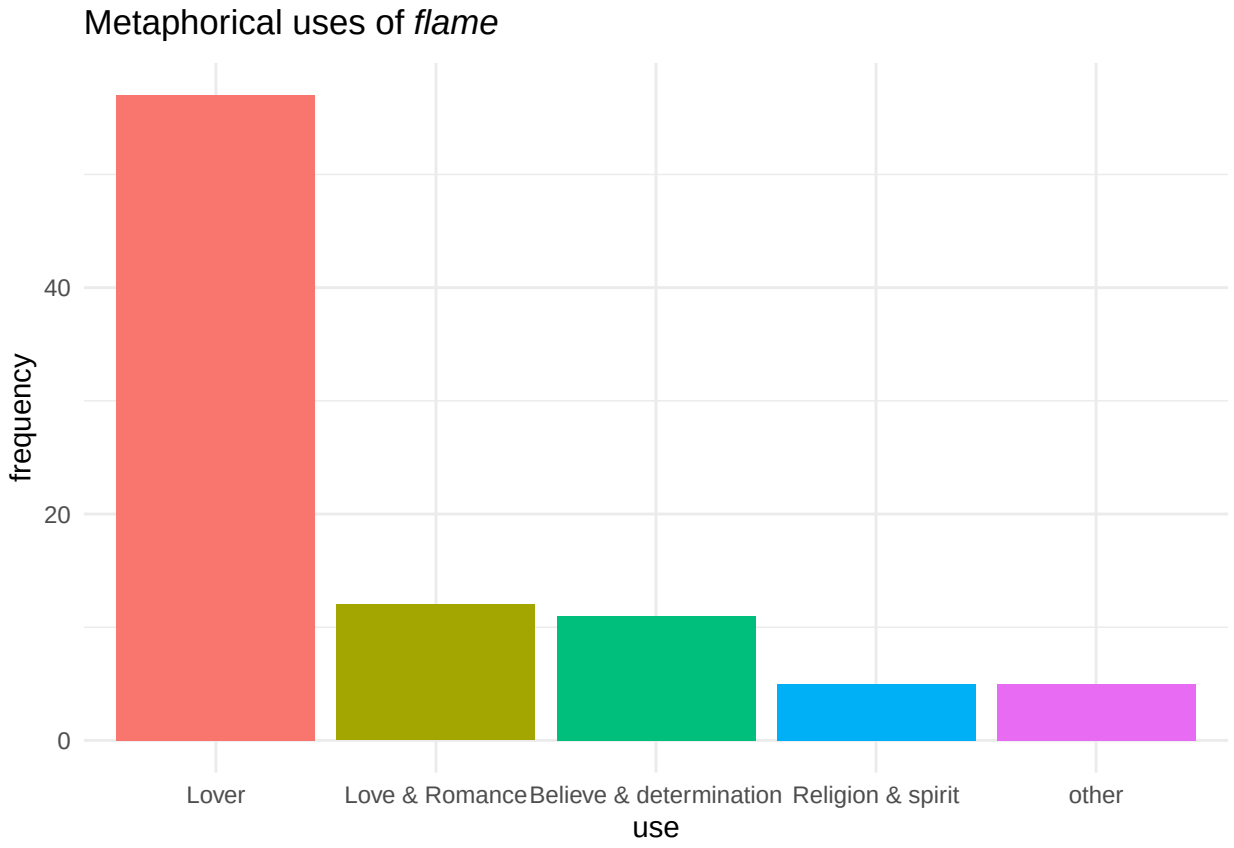
- Absolute frequency
  - Basic measure
  - Should always be reported since everything else is based on it
  - Sometimes hard to visualize
  - Hard to interpret across different sample or category sizes
- Relative frequency
  - Absolute frequency divided by all occurrences
  - Either between 0 and 1 or 0% and 100%
  - Makes it possible to compare between different sized samples or sub-categories
  - extremely low relative frequency is sometimes reported as normalized frequency, e.g. 1 per Million vs. 0.000001 vs. 0.0001%

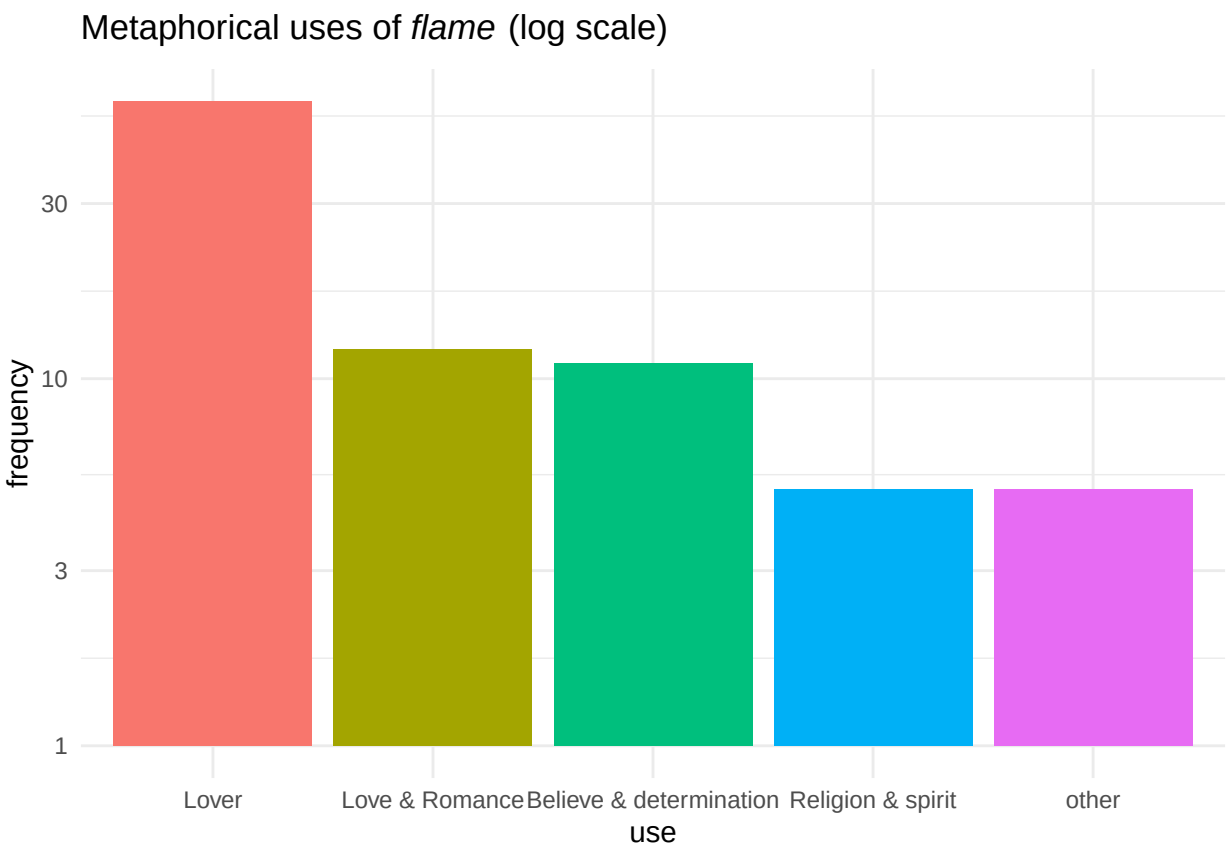
6.1.4 Frequency and scales

- Log scale
  - Most commonly base 10, i.e. 1 to 10 is the same distance as 10 to 100, 100 to 1,000, etc.
  - Uses:
    1. Visualize heavily skewed data
    2. Make exponential data linear (e.g. word counts)
    3. Approximate human perception of quantities
- Pie charts
  - variant of stacked bar chart
  - becomes hard to interpret with many categories

6.1.5 How extreme are the differences?

If we take into account the proportional nature of our perception, we might get a better picture of frequency differences on a log scale, which essentially emphasizes proportional differences rather than absolute ones. If you compare the two graphs below you might not change your conclusion about the data, but the felt difference between the categories might be much smaller than you would think. As stated above, these considerations are to be taken with a grain of salt on such small data sets.





6.2 Metaphor and Cognition

Metaphor is not merely a figure of speech that we can use to write nicer poems or prose. It is so fundamental to our perception of our environment that it is in fact our main way to construct non-perceptual information.

6.2.1 Time is space

One of the most pervasive metaphorical patterns in language is that of *time as space*. You might easily overlook this mapping because understanding temporal relations in the sense of spatial relations is so basic to our cognition that it is easily overlooked. We do not have a physical sense of time. Our eyes provide us with information about space; depth perception allows us to sense differences in distance. We can also directly perceive ourselves relative to the space we move in with a combination of our senses of balance and proprioception. All the information of time is inferred from those more basic senses and the changes we experience. Language use reflects this asymmetry.

Historically, many, if not most, of our temporal function words are etymologically derived from spatial function words or lexemes with a spatial meaning.

- Prepositions: *at, after, before, between* …
- Temporal auxiliaries: *going to, be about to, venire de* (fr.), *voll* ***am*** *Chillen, Digga* (ger.) …
- Temporal adverbs: *always, next* …
- Nouns: *presence, past* …
- …

6.2.2 Exploring color metaphors

I asked you to brainstorm color metaphors and you split up in groups. You found a lot of interesting mappings and linguistic structures representing them. Here are some highlights:

- Red as danger: *red light, red flag, redlining*
- Red associated to communism: *red army, red menace*

You also found that red is associated to beauty, love and sensuality. This association is mostly related to things that are literally red so we couldn't find many clear metaphorical structures. There is the *red light* district, though.

- Blue is calm: *feeling blue, out of the blue*,
- Black is bad: *black night, black death, black times*
- Black is obscure, unnormal: *black market, black money, black sheep*
- Green is young or unexperienced: *green boy, green behind the ears*
- Green as environmentally conscious: *green politician, green(er) cars*
- Green is jealous: *green with envy*

The Yellow group came to the conclusion that yellow seems to be used mostly literally. You hypothesized that the color is more important in Eastern cultures so we could expect to find more collocations and idioms with yellow used metaphorically.

6.3 Homework

Next week's topic is Metonymy. Please read Deignan (2005). There is not going to be a regular live session on Monday. Instead I am offering you two choices. Either do a homework assignment with me in small groups in a live session or do it on your own and send me a little summary. Here is what we'll do.

Brand names: We will look at brand names such as Microsoft, Apple and Google and analyse their use with a focus on metonymy. The aim is to explore, query and download corpus data for further annotation. Which brands exactly we will decide together.

There will be four sessions from Tuesday to Friday each between 16:00 and 18:00. I'd say no more than 6 people per group. I've prepared 4 channels on Discord for every day. Just write "dibs" in the channel for the day you'd like to join.

This is an offer since interaction with such a large group is difficult. So use it! :) For those who decide to do it on their own or did not catch a spot, there will be specific instructions on the assignment next week.

6.3.1 Tip of the day

Use spreadsheets! I encouraged you to create simple graphs in the last homework. That required that you enter some numbers into something like *Excel*, *Calc* or *Google Sheets*. We will benefit from spreadsheets throughout this course, but this is not where their utility stops. Being able to do some quick formulae and vlookups in Excel are common skills used outside Uni.

Especially for teachers, I believe, spreadsheets are an essential skill: for grades, averages, homework, quick stats on exams, lesson planning, Sitzplan (oh memories :D), what have you. If you know your way around Excel, you can speed up your tax returns (Steuererklärung) a lot, too. Many teachers end up working as freelancers. For a freelancer (and anyone else

really), gathering your receipts, bills and pay slips neatly arranged and categorized as data in a spreadsheet can save you endless amounts of time and even money.

This is not where it stops though.Timetables and To-Do-Lists are also neat to do in a spreadsheet if you need more fine-grained control over the layout than the clunky online calendar you are probably using. Here are some things I have used spreadsheets for in the past: notes, training log, travel plans, shopping lists. You could even use them for recipes or counting calories if that's what you're into. I've since moved past Excel and use only plain text files. If I need to do some maths or stats I use .csv and R. That would be the ultra-nerd level so don't be scared of a spreadsheet ;).

7 Metonymy

Metonymy is a is a concept closely related to metaphor and another way to extend the use of a lexical item, and therefore, an important structure. Deignan (2005) argues that metonymy and metaphor are two sides of the same coin, and it is, in fact, difficult to draw a line sometimes. Typical examples include the part-whole- and whole-part- relationships. We can refer to smart

(23) The university had sacked Mr Jeffries. (BNC)
(24) Then I had to come back and read Shakespeare (BNC)
(25) Apple announced there new app on Monday.

7.1 Group 1: Exploration, or 'get off my Amazon'

We decided to look for metonymical uses of *Amazon*. The corpus data needs to be quite recent since Amazon wasn't called Amazon before 1995, and the web services which made it famous were started only in 2002. Many of the popular corpora that are large and balanced, like the BNC, take a lot of time and resources to be compiled. As a result, they are sometimes too old for some types of research question. While the BNC is perfectly fine for a large range of grammatical and lexical topics, it is too old to show the company name Amazon.

```
[no corpus]> BNC
BNC> "Amazon"


…
 1548472: hern Colombia , especially its [[[ Amazon ]]] cocaine laboratories on the b
 2187606:  , a town in the jungle of the [[[ Amazon ]]] Valley . Having read Schiller
 3307845: een of the lower waters of the [[[ Amazon ]]] River . When they were wet fr
 3317618: red miles from land the fierce [[[ Amazon ]]] river stained the dark water
 4197985: the next ten years much of the [[[ Amazon ]]] Rainforest could be wiped out
 4198128:  the systematic burning of the [[[ Amazon ]]] Rainforest . TOMORROW WILL BE
 4198254: at current rates , much of the [[[ Amazon ]]] Rainforest will have been obl
:
```

All hits are related to the river/region. This is an extreme example, but you always have to make sure that your corpus is representative as a data source. The results of corpus queries are vastly dependent on the makeup of the corpus. The more automatic your data retrieval, e.g. relying on available annotations and frequency lists, the more dangerous unexpected patterns or the (unexpected) absence of expected patterns might become. For instance, if you were looking for a whole class of names of which *Amazon* is only one, aspects like this might not be immediately be apparent. Automatizing coding in linguistics is very powerful, but with great power comes great responsibility.

You can get information on the corpus by typing `info`. There, you can find available attributes (is the corpus lemmatized, pos-tagged?), textual annotations (mode, genre, author/speaker),

and general information. If the information in the info file is not enough, note that there is usually a publication connected to a corpus, which is the one you also have to cite if you use it as data source. For example, if I use the Corpus of Contemporary American English (COCA), I cite Davies (2008).

So, for *Amazon*, we need a more recent corpus. A popular choice is newspaper corpora, which can be very large and very up to date. A major disadvantage is that they are only representative of newspaper language. There are also problems with copyrights and paywalls with many corpora. We do offer some newspaper corpora, and there are some available online. For now, let's compromise on a rather recent corpus that is also reasonably large. We have the spoken version of the new BNC 2014, which you can activate by typing `BNC2014-S`.

In the 2014 spoken data, we are more lucky. In fact, most of the matches appear to be about the company rather than the place. In order to get rid of the forests and rivers, we could try to look for patterns that only occur with the geographical name and don't occur with the company name. We noticed that most occurrences are preceded by *the*. In order to see whether we can exclude them systematically, we first looked at all of those matches.

```
BNC2014-S> "the" "Amazon"

…
 763807:  of money to do it and that  s [[[ the Amazon ]]] to the Andes oh nice erm but
 766016: can see him going through like [[[ the Amazon ]]] and stuff and they get like r
 790850: the and like river dolphins in [[[ the Amazon ]]] you get you get like river do
1537836: e okay the Himalayas it  s got [[[ the Amazon ]]] it  s got everything yeah and
1694346: r Kindle right because mine is [[[ the Amazon ]]] Kindle then that  s where it
3551760: en yeah and he he travelled in [[[ the Amazon ]]] he followed the river for fif
3773906: osite side of the mountains is [[[ the Amazon ]]] in pretty much like all of La
3774293:   s but part of that is in the [[[ the Amazon ]]] yeah and he  s he  s gone dow
3774587: I mean I did n t spend long in [[[ the Amazon ]]] and it  s hard work mm it s
3774745: ver fainted was when we got to [[[ the Amazon ]]] really ? yeah it was quite sc
```

This filter looks rather successful, but we do get the company name when it occurs in certain attributive uses, such as *Amazon Kindle* or *Amazon delivery*. The number of matches is small enough to actually clean it up manually, but in a larger sample you would want to optimize your query more. For now, we were ok with the results. We might want to exclude attributive uses anyway eventually since they are rather different from the other nominal uses.

To exclude the results rather then restrict to them, we use the `!` operator which is a logical *not*: `[word != "the" %c] [word = "Amazon"]`. Note that the bracket notation `[word = "word"]` is the same as using the shortcut `"word"`. This should be your general approach before you exclude anything. Check what is in there before!

An interesting and maybe unexpected pattern that we found while browsing through the data is the fact that Amazon is used metonymically to mean the Amazon online account. This is the same that happens when your parents ask you to send them *a whatsapp*, or when people are looking for a *Kleenex*.

(26) I have stuff on my Amazon as well
(27) can you get off my Amazon?
(28) I'll go on to my Amazon.

Characteristic of this use is the fact that we use possessive determiners *my, your, their, her, his* in front of them with no noun following (remember: attributive uses). We can figure out how to search for possessive determiners by looking up the right tag in the info file typing `info` and searching with `/` for "possessive". The relevant tag is `APPGE` so we can search for this by using `[pos = "APPGE"]`. We might want to identify other possessive structures like genitive *'s* and consider other structures that are characteristic of this use.

The results on *Amazon* in this corpus are rather limited, but the `possessive + brand name` structure gives us a nice place to start looking into the Whatsapp-Kleenex-Tempo type of metonymy.

7.2 Group 2 and 3, or 'comparing Apples to Mangos'

In group 2, we started off with the same logic as outlined above, so I am not going to repeat everything. Instead of *Amazon*, we looked for *Apple*.

One aspect worth mentioning about querying in CQP is that everything is interpreted literally. That means that it makes a difference whether we search for `"apple"` or `"Apple"`. When it comes to proper names this can be helpful and get rid of many false positives referring to fruit. If you don't want this behaviour and want to include all permutations of capitalizations, you can append `%c` to the end of every token or after word when you `count`. This literally means "ignore case".

In a nutshell, to improve our results, we excluded *Big* as in *Big Apple* and we decided to exclude any attributive use. To achieve that, we excluded any pos tag that starts with *N*, using a regular expression: `[word != "Big"] [word = "Apple"] [pos != "N.*"]`

Among the results, we spotted some metonymies, most of which included personification. In order to explore those personification contexts more, we restricted to any occurrence directly followed by a verb. This gives us mostly subject uses of *Apple*.

```
[word != "Big"] [word = "Apple"] [class = "VERB"]
```

Bare in mind that those queries are not exact and only for exploration. `NOUN + VERB` is not a sufficient query to find subjects reliably, much less personifications. But these some steps you can take to begin to filter your results.

As a next step, we widened our scope and included more and more related brands into our query by stringing them together with the logical *or* `|`. We were trying to define a list of of Social media brands.

```
[word = "Apple|Microsoft|Google|Facebook|Twitter|Instagram|ICQ|Windows"]
```

In a real study this list should not be arbitrary and best be exhaustive, meaning you should include all brands that match certain criteria you define first.

In group 3, we took the above results and made a comparison with fashion brands. While tech and social media brands seem to frequently occur in personification contexts, we could not find the same behaviour in fashion. Rather, we found that the metonymies seem to be mostly in combination with local prepositions. In order to have a first test on this impression, we can use the `count` command.

```
[word = "Adidas|Place|Levi|Nike|Primark|H&M|Lacoste"]

`count by pos on match[-1]`
`count by class on match[1]`
```

32 out of 106 matches are preceded by prepositions while only 14 are followed by a verb. As a next step we would have to make our lists of brands more exhaustive, our queries for both categories more robust, and compare the co-occurrence frequencies properly.

The tentative hypothesis we drew from this short exploration was that Social Media brands are conceptualized as humans/actors while fashion brands are conceptualized as places, which is quite exciting for 90 minutes of playing with data.

7.3 Group 4, or 'How many are Greenpeace anyway?'

In the last group, we basically used all of the techniques above. One major difference was that our object of interest, *Greenpeace*, was old enough to study in a larger corpus, in this case the original BNC.

We discovered that Greenpeace is used both as a collective noun, i.e. a noun that can both act as a plural or as a singular syntactically.

(29) That is why Greenpeace have had to take the moral initiative.
(30) but its Greenpeace has 600,000 members

There is a variety of interesting questions that we could ask at this point. When is there plural agreement and when is there singular agreement? We could also see whether the association has an influence on the choice. Is *Greenpeace* like any collective noun, such as *police* or *audience* or are there differences? What is the function and what is the effect of using either plural or singular? What is the relationship to metonymy in this case?

8 Synonymy

> It's passed on! This parrot is no more! It has ceased to be! It's expired and gone to meet its maker! This is a late parrot! It's a stiff! Bereft of life, it rests in peace! If you hadn't nailed it to the perch, it would be pushing up the daisies! It's rung down the curtain and joined the choir invisible. This is an ex-parrot!"
>
> — Monty Python[10]

8.1 The same meaning and function

Synonyms are commonly understood as existing on word level, or lexeme level. As illustrated in the quote above, it is possible to have phrases being synonymous with words, phrases with phrases, etc. In fact, any lexical entry qualifies, i.e. we could also meaningfully describe synonymous affixes, function words or even entire constructions or syntactic patterns. The essential property of a set of synonyms is that they have the same meaning. In modern usage-based theories, such as Cognitive Linguistics and Construction Grammar, the lines between meaning and function get blurred. One main idea is that the way we use words and the contexts they are found in defines their meaning.

Those frameworks do not assume an objective reality and meanings based on truth conditions and referents that exist outside language. Instead, meanings are understood in terms of **construal**, that is the way reality is perceived subjectively. The idea is that we don't call a book a *book* because there is a group of objects that exists independently from language, but because there are objects in our experience that we perceive of as *book*. This might not sound like a very useful description, but it elegantly captures meanings which are difficult to anchor in the "real world". Just think of unicorns, orcs or ghosts. We can communicate information about those creatures even though we lack hard evidence for their existence. We might even disagree whether there is an actual referent in the real world, as is the case for ghosts. What's more, our communication about them isn't any different from that of "real" objects. *unicorn* is a noun like *horse* with the same set of affixes, obeying the same syntactic rules. If we base meaning in human experience rather than objective truths about reality, we can explain lexemes describing fantastic creatures without having to assume multiple realities, which would have been the traditional approach.

[10] http://www.ibras.dk/montypython/justthewords.htm

If meaning is bound to experience, and we use language to construe those experiences, meanings become defined by the context of use. In some sense, the way we use a word defines its meaning. In order to explain synonyms, we have to assume that they have the same meaning, i.e. the same, or almost the same, use.

8.2 Principle of no synonymy

If we blur the line between meaning and use, a consequence is that true synonyms would have to occur in the same lexical, grammatical and discursive contexts. Corpus linguists have been quite successful, however, to show that this is not the case. Most synonyms aren't interchangeable at all.

(31) We made big plans for 2020.
(32) ? We made large plans for 2020.

Some synonyms would create unidiotmatic utterances, i.e. utterances that sound very unusual to a native speaker.

(33) Let's have a drink tonight.
(34) ? Let's have a beverage tonight.

Even synonyms that seem to have contexts in which they are interchangeable show meaningful patterns. You can predict when people use *start* vs. *begin* probabilistically, based on grammatical and lexical properties (Schmid 1996; Divjak & Gries 2009). E.g. it is much more likely to find *begin to do* than *begin doing*, even though you can find both (Schmid 1996).

Among other things, the observation that there are subtle, but non-random systematic differences between even the closest of synonyms have led some usage-based linguists to that there are actually no true synonyms at all, but only partial ones.

Goldberg's (1995) principle of no synonymy:

- If there is a difference in form, there must be a difference in function.
- There are no **true** synonyms.
- Synonyms are used in different grammatical, lexical, discursive contexts.
- Synonyms may be loaded with dialectal, socio-cultural associations, which are not inherently different from other grammatical and lexical properties

Why are the following utterances unusual?

(35) Daddy! The chocolate is so nice! We must **purchase** more! (anecdotal)
(36) The perpetrator left in order to micturate. (Elementary, TV series)

purchase is found mostly in contexts of trade, acquiring larger amounts of goods, legal language, or in general higher registers and when the process of buying itself is in focus. A child using it in a colloquial context for something like chocolate is, therefore, unexpected and sounds unusual or marked. The other example is uttered by Sherlock Holmes, who is very aware of his intellectual superiority and who might want to distinguish himself from other people by using precise and scientific language. Using synonyms provide a way for social distancing in a conversation. The opposite is also possible and can be frequently observed in something that is called the chameleon effect. People who speak a dialect might change their language substantially when they are around people that speak a different dialect or adhere to a standard. Often, the dialect speaker will switch to the standard at work but speak dialect at home with their parents or friends. Word choice is a salient example for the chameleon effect in action, but the same mechanisms are true on all levels of language, from phonology, over grammar to discourse.

8.2.1 Dative alternation

An example for a pair of construction that is very often thought of as synonymous is the so-called dative alternation.

(37) Double object: *I gave you the tickets.*
(38) To-Dative: *I gave the tickets to you.*

There are categorical contexts in which the constructions are not interchangeable. E.g. you cannot use a pronoun in the as second object in a double object construction with a heavy first object (42), i.e. a long or complex first object.

(39) *I gave it to you.*
(40) ? *I gave you it.*
(41) *I gave it to my best friend in the world.*
(42) * *I gave my best friend in the world it.*

What's left is non-categorical contexts where we do find a substantial amount of variation.

(43) *I gave my mother a very nice present.*
(44) *I gave a very nice present to my mother.*

8.3 Homework

I am planning to make a series of casual live streams during which I am going to work on one of the research questions that came out of our group sessions last week. This could give you a better idea of what goes into working on a linguistic project in practice. I will keep it real, use my own workflow and programs, but talk you through my line of thought.

You can vote on the topic and on the day I should stream. Participation in these streams is totally voluntary. You can drop in and out whenever you want. I will start at a scheduled time and then see what happens. There are only three topics because I collapsed ideas from Tuesday and Wednesday since they were quite similar.

Depending on the interaction and feedback from you, this might go as far as resulting in an example term paper. Who knows. It's gonna be an experiment.

Here are the links to the polls. You can choose multiple options:

- Topic
- Time

Check the summary of week 7 for how we got to these ideas during our group work.

8.3.1 Tip of the Day

Today: Multitasking

Learning an academic discipline takes a lot of time and focus. However, some aspects are like learning a language or motor skills. It might sound weird, but knowledge, especially theoretical, is like a muscle you can train. So here is my suggestion for how to get better at Linguistics or Literary Studies or whatever science you are interested in: Listen to lectures, talks, podcasts and other content in the background.

Great topics to passively consume are:

- Theory, e.g. Cognitive Linguistics
- Philosophy of Science, highly interesting, vastly important, but oft neglected
- Sciences that are not your major

Here are some activities I frequently use to bombard myself with knowledge.

- weight or endurance training
- practicing an instrument (especially repetitive technical exercises)
- cooking
- cleaning, tidying, building Ikea tables ;)

Non of these activities require your full mental focus or have long pauses, so your thoughts are free to meander through the depths of science. Nowadays, a lot of talks or even full lectures can be found online, and with online teaching taking off right now there will be ever more.

Linguistics Luckily, we are not the only university trying to teach you linguistics online. Here are some nice channels to binge watch both actively and passively.

- Martin Hilpert: Has a variety of lectures and full courses on all things linguistics.
- The Virtual Linguistics Campus: Old but gold.
- People without YouTube channels, but who are great lecturers, Adele Goldberg, Joan Bybee, George Lakoff, Geoffry Pullum. I have found many of their lectures and interviews online on various channels and platforms.
- NativLang: Probably my favorite language channel. Animation videos on a variety of language related topics. Focus on Cross-Linguistics.

Other sciences If you are a curious person, and if you appreciate the academic endeavor, chances are you are interested in other sciences, too. Knowing subjects outside the social sciences may help you in unexpected ways. Here are my go-to channels to listen to in the background.

- mailab: Focus on (bio-)chemistry, but mostly deals with current debates on the media. You can learn a lot about how news outlets interpret and sometimes misrepresent scientific studies.
- PBS Space Time: Astrophysics. Popular science without the usual dumbing down. Great stuff to listen to even if you understand nothing. :D
- Closer to the Truth: Philosophy. Dealing with the big questions. How do we know facts? Why should we trust in Science? What are hypotheses and theories and why bother?
- Statquest: Pleasantly cringey statistics videos.
- zedstatistics: More in depth. (Less cringe. :( )
- 3Blue1Brown: Mathematical concepts with animations instead of formulae. I was horrible at maths in school but I always had a sense that it is actually a very beautiful subject. Wish I had visualizations like these back then.
- Computerphile: Various computer science topics

I have not yet explored the world of audio books and audio podcasts, but I'm sure there is a lot of great stuff out there.

If you discover anything, let me know! :)

9 Antonymy

9.1 How do antonyms emerge?

Last week, we looked at lexical items that have the same meaning—synonyms. We had a first exploration of the implications of Cognitive Linguistic on meaning-related phenomena. We got to know the Principle of no synonymy, which claims that there must be a difference in meaning if there is a difference in form. There is a strong focus on usage patterns in this view. Meaning is seen as being inseparable from use, therefore, co-occurrence patterns become extremely important.

If lexemes cannot have exactly the same meaning, can they even have opposite meanings? Does it even make sense to speak of opposite use? If we reject a componential model of meaning, antonyms become a problem at a first glance. There are two main ways to explain antonymy. The traditional one is that antonyms can paradigmatically replace their opposite.

Paradigmatic, replaceable

(45) He was a good dog.
(46) He was a bad dog.
(47) I feel good today.
(48) I feel bad today.

In fact, that makes them extremely similar in their use. You would expect similar collocates, constructions and syntax.

Justeson & Katz (1991) argue against this view and propose that the intuition we have that some words have direct opposites is grounded in their co-occurrence syntagmatically.

Syntagmatic, co-occurrence

(49) There are good and bad dogs.
(50) Some dogs are good, some are bad.
(51) I feel neither good nor bad.
(52) Good jokes make people laugh, unlike bad ones.

One fascinating aspect of these observations is that, while synonyms occur in wildly different contexts, antonyms tend to occur together. The hypothesis is that we think of antonyms as antonyms exactly because we see experience them together all the time. *high* and *low* are contiguous; they co-occur—*high* and *flat* are not, because they don't not because of some objective meaning components. In fact, there is usually only one antonym within a set of synonyms. If we assume a componential model based on truth-conditions, this would be difficult to explain. Taking logical meaning components alone does not explain speaker intuition.

9.2 Causal relationships

The findings from Justeson & Katz (1991) work very well with cognitive concepts of memory and learning. An interesting interpretation of the findings would be that we don't need any inherent meaning to explain antonyms. Children would just learn what antonyms are through language use. Here, we are falling for a common trap though, which is bias towards a specific theory. The findings might be consistent with more than one theory. In order to evaluate the usage-based interpretation, we should also consider alternative explanations.

Usage-based linguistics is most strongly contradicted by nativist theories, such as Universal Grammar. The idea of nativists is that we are born with a capacity for language including some deep undelying linguistic categories. In this view, children already have categories like antonymy (or more generally oppositeness) hard-wired in their brains. Language learning then would consist of categorizing new stimuli against these pre-existing categories. It would be possible to imagine that the structure for antonym relationship is already given and children learn which words are antonyms, and as a result of that use them together more often than other word pairs.

Ultimately, it is a question of causality. Did co-occurrence cause the emergence of antonym pairs, or did the oppositeness of the lexemes cause the co-occurrence? We have arrived at a chicken or egg situation:

We discussed these two interpretations in groups:

1. Antonyms co-occur and children/learners associate them. Their oppositeness is a result of this.

2. Antonyms have opposite meanings, children learn that first and, as a result, use them together.

Our discussion went very deep very quickly. One of the main points were the search for an a priori definition of antonymy or oppositeness that we would need to determine a potential pre-language concept. Observational methods are not normally used to infer causal relationships. This is normally the job of experimental methods, where you can control for confounding variables. However, when it comes to children, the options are limited and it is hardly possible to check whether there is a pre-language concept of antonymy. Without empirical data, the options are limited but, you can still theorize over the most likely explanations. Non-empirical methods are common in literary and cultural studies and also philosophy.

The other major point was anecdotal evidence. When it comes to language and children, there have been a number of stories and questionable unethical experiments. Friedrich II. and the Nazis experimented on children and deprived them of language, and there have been multiple accounts of orphaned children who grew up alone or with animals. It is tempting to take these stories as evidence. However, neither of these accounts have been carried out with the necessary academic rigor, and there was usually some sort of ideological agenda. Anecdotal evidence is often full of contradictions and fantasy. Sometimes the stories are distorted to fit a particular belief, and sometimes contradicting aspects are simply left out. For the most part, observations that are non-reproducible are not viable even if (or maybe especially when) the reasons for their non-reproducibility are ethical.

In conclusion, many observations in corpus linguistics are simply evidence for correlations, and it is very hard to infer causal relationships without the help of experimental methods. If those are not available, it is necessary to have a good grasp on the philosophy of science to narrow down possible interpretations. In any case, you always have to be aware of the limitations of the data and methodology.

10 Lexical Patterns

10.1 Multi word patterns

Over the past weeks, we have been focussing mostly on properties of individual lexemes. Of course, there was always an emphasis on the fact that lexical entries are not restricted to word level. We have encountered compounds and complex prepositions which, in some situation, behave like other single-word lexemes. However, we haven't really focussed at the syntagmatic relationships between lexemes. In grammar, we would transition from morphology to syntax now. What we could see as equivalent in lexical semantics, is multi-word patterns, which range from simple adjective-nouns collocations, to clause-sized idioms.

10.1.1 From collocations to syntactic patterns

(53) tall building
(54) remarkable feat
(55) stark contrast
(56) thick accent

With collocations, we have encountered a common class of multi-word pattern already. Most of you should also be familiar with a range of syntactic patterns, such as Subject-Verb-Object, relative clauses, the double-object construction, tenses etc. Phrase structures (remember syntax trees) are inherently multi-word patterns. Many of these structures have aspects of multi-word patterns, and often include lexical elements. Tenses, such as the going-to future (57, 58) have a fixed element (e.g. the semi-modal going to/gonna) and a variable verb phrase slot. Not every lexical verb can be used with all tenses. As a result, you get the patterns that every learner

of English is familiar with. For example, stative verbs don't usually occur in the progressive (60). Any syntactic structure has a lexical component to it.

(57) They are going to open the gyms again.
(58) They're gonna open the gyms again.
(59) I am eating at the moment.
(60) * I am knowing it at the moment, but I might forget.

going to is an interesting case because of the direction of the contraction to *gonna*. With phrase structure as the only model to explain the syntagmatic relationships, we would have to expect the particle *to* to be contracted to its closest constituent, which would be the following verb, with which it forms a unit (to-infinitive). What we actually see is *going* and *to* fusing together. This is easily explained just by looking at frequencies. Imagine that our brain works like a sort of computer that constantly tries to model and predict incoming stimuli and automate and reduce wherever possible. The probability that *going* is followed by *to* is very high, in certain uses you could consider it redundant information. The verb slot following *to*, however, is much more variable. Therefore, even the most frequent combinations of to + verb aren't frequent and distinctive enough to trigger contraction.[11]

10.1.2 Lexical variability

If we think of lexical patterns and syntactic patterns as two sides of the same coin, we still have to account for their obvious differences. Syntactic patterns are much more general and have optional slots and highly variable slots. Collocations on the other hand are sometimes part of larger variable patterns, but there is always some lexically ivariant part. Below we have elements like *ear off* and *into* that don't change. The other slots in those constructions, however, do. Nevertheless, the variable slots cannot be arbitrarily filled. The verb slot in (61) is largely restricted to communicative verbs, in (62) the noun slot only allows a very restricted set of swear words.

(61) [VERB] [SOMEONE]'s ear off
(62) [beat|kick|slap] the [crap|hell|shit] out of [SOMEONE]
(63) into-causative: [VERB] [SOMEONE] into [VERB]ing [something]

In conclusion, the most general and most variable structures would be found on the side of syntactic patterns, while the more fixed constructions are determined lexically. Lexical and syntactic patterns co-exist and readily mix and mingle.

10.2 Multiple levels of generalization

There being multi-word patterns that cross phrasal boundaries and contradict traditional grammar rules does not mean that the old models were wrong. English still has a hierarchical phrase structure and an SVO word order. What we can see, however, is that categories are formed on different levels of abstractions. Syntactic patterns, such as word order, tense and sub-ordination are much broader generalizations. Those structures reliably show up in creative language use. Collocations and most other lexical patterns are more idiomatic, bound to certain lexemes.

10.2.1 Competing motivations

There are several explanations of why we would have competing systems. One such explanation is that there are actually competing motivations shaping language.

[11] Of course, this is a strongly simplified explanation of how contractions happen. The exact mechanics behind grammaticalization processes responsible for things like the *going-to*-future are a topic in Historical Linguistics.

Automation:

On the one hand, communication systems are under pressure to be more time and energy efficient. We see trends in language that make frequent utterances shorter and more specific.

Idiomatic language is more time and energy efficient in highly specific recurring situations

- Utterances are more easily processed when they are **strongly associated** to specific communicative situations
- Specific lexical structures avoid ambiguities
- Fine-grained conventions allow for stronger group identity

Simplification is more efficient in constantly changing, unpredictable contexts

- Broad generalizations allow for creative language use
- Humans need to categorize the complexity of stimuli
- Unspecific linguistic items are easier to use across speech communities

10.3 Homework

Homework for you today is just a bit of practice with the system.

1. Pick 2 of the bundles discussed in the text and search for them in a corpus of your choice.
2. Create a frequency list of the most commonly occurring words to the right of your bundle using `count`.
3. Create a concordance of the other bundle with 4 tokens to the left and right.
4. Export both as a .csv file, download them and import them into a spreadsheet program.

Don't worry if you don't manage to do all of the steps. Just let me know how far you got. Feel free to use the Discord chat as support forum.

5. There are two ongoing evaluations. The first is from our linguistics work group that we're using as a guide for the coming semester. The other one is the regular evaluation. We would be very grateful if you could take a couple of minutes and participate in them. Especially the first one has already been very helpful.

- Our work group's evaluation: Click here
- Official evaluation: Click here

Check out the links section. I have uploaded a repository with the project I am working on during the term paper live streams. You can fully track my entire progress and use it as inspiration of even as a template. There is a new link to a page where you can provide anonymous feedback and complaints.

10.3.1 Tiwilbemba

Today I am sharing with you my biggest regret looking back on uni days, thus, my biggest tiwilbemba: Not learning LaTeX/Markdown early enough.

Many of you are no fans of sitting in front of the computer all day. If you are a student, you will use a significant amount of time writing essays, term papers, and theses. The biggest time sinks with these are formatting, tables of contents, bibliographies, lists of abbreviations, etc. What if I told you that you don't have to spend any time with this? If you know just enough LaTeX/Markdown, you can skip over all these steps, which equals less time tinkering at the computer. If you watched my first term paper stream, you literally saw me set up a document from scratch in under 5 minutes, including cover sheet, table of contents and bibliography, everything formatted perfectly and updated dynamically as I fill it.

It might feel counter-intuitive to spend even more time learning an entirely new computer skill. But bear with me. The time you spend on learning how to write documents in LaTeX or

Markdown is ridiculously small compared to the days if not weeks of formatting frustration you can save yourself. I have always been rather tech savvy, and I know Microsoft Word much better than, I guess, the average user. Still, in hindsight, I feel like I was wasting my time. I wrote all my seminar papers, essays and theses in Microsoft Word. And I regret it.

This section is not a tutorial, rather an encouragement for you to expand your horizon (even though I will upload a simple set up for a term paper in the appendix soon). First, a profile of people that should, in my opinion, learn writing in plain text (LaTeX or Markdown).

Group 1: You have to write…

1. Academic papers
2. Reports
3. Articles
4. Books

Anything that requires a simple style that doesn't require a crazy amount of design greatly profits from LaTeX/markdown. Any repetitive work that requires consistent formating, too. If you write larger works like books, you'd be crazy not to use LaTeX. Students definitely belong in this group. I'd say, if you force yourself to learn it now, by the time you write your bachelor thesis, it will have been worth it already.

Of course, there are people who might be happy with graphical programs. To be fair, let's profile these people, too.

Group 2: You have to write

1. not much at all, only the occasional document
2. Constantly changing documents
3. Design-heavy documents (e.g. Ad material)

If you belong to this group, you might not profit from learning LaTeX too much, and you probably don't care for Markdown either. Creative design is difficult, unless you are very experienced already.

Here are some reasons people have against learning LaTeX that are not valid in my opinion.

1. "It's difficult."
   As soon as you've set it up and learned the basics it is actually sooo much easier. There are also platforms with great communities like Stackoverflow, where almost any problem you encounter has been solved by users with full examples. You just have to search for it.
2. "I am not a programmer"
   Neither am I. Don't let the syntax scare you.
3. "I'll learn it eventually, but for now I have to get this paper done quickly."
   Nope… That's what I told myself up until a year or so ago.
4. "I'll need to work with people that use .docx."
   A good .tex file can be easily transformed into .docx or .odt thanks to tools like Pandoc.

Finally, some reasons people might not consider normally.

- *Professionals love it*: If you were to write a program, you'd ask a programmer how to do it best. If you were to build a door, you'd ask a carpenter. For some reason, if people do typesetting, they do not use the tools of professionals. Most publishers use LaTeX, and also accept Latex files. It's definitely not a bad thing to put on your résumé either.
- *Focus*: not seeing the output immediately is actually a great thing. You might have just hopped onto the train of thought and the words just spill onto the screen when,… Hark! The table you placed so carefully a moment ago moved unexpectedly to the wrong page… Moment over, distraction has won. This is not gonna happen with LaTeX/markdown. I personally find myself micromanaging all the time in word.

- *Light weight*: if you have an old computer or laptop that is old or cheap (or pretty, expensive but still weak,...you know) Windows and Microsoft Word/MacOS and pages might actually run rather slowly. If you think they are fast, you haven't experienced the alternative. Especially large documents might take some time to load. If you have everything in plain text files, you're document loads in a split second. That might take away some subconscious blocks that prevent you even from even opening your project. Just pop it open and quickly add a thought to your paper. Sooo comfy. :) As a matter of fact, I'm currently writing this very article from my phone using an editor called Markor with my source file synced in my cloud.
- *Gateway drug*: Writing your term paper in plain text might just be the beginning. If you understand LaTeX, you basically get html for free. The principle is the same, just with slightly different syntax. If you use something like Rmarkdown, you can essentially export your project seamlessly into any format with little adjustment needed. You might be tempted to write your own website. Maybe you get into extensible text editors, terminals, scripting, maybe even Linux, maybe even... Vim? The rabbit hole goes deep. ;)

11 Lexical Bundles

Coming soon...

11.1 Prefabricated Chunks

11.1.1 Examples from the seminar

also prefabs, lexical chunks, lexical bundles, ...

(64) *one of the most*
(65) *I just wanted to*
(66) *I was gonna*

(67) *the fact that the*

11.1.2 Prefabs and the lexicon

Automation

- phonetic reduction (cf. also Bybee & Scheibman (1999))

Constituency

- pauses tend occur after frequent chunks rather than at syntactic boundaries
- are learned, and accessed as units

Source for lexicalization and grammaticalization

- Discourse markers
- Tense and aspect markers
-

11.1.3 Discourse markers

A large subset of linguistic structures exist for purposes other than exchanging information

(68) *on the one hand*
(69) *after all*
(70) *for the first time*
(71) *nevertheless*

11.1.4 Common Functions

- turn taking:

(72) …, right? (turn releasing)
(73) *yeah right but* (turn yielding)
(74) *uhm well, you know* (turn holding)
(75) *yes*, *oh my*, (signalling attention)

- establishing a common point of view

(76) *you know what I'm saying*
(77) amirite?

- deixis

(78) *therefore*
(79) *in the last section*
(80) *in conclusion*
(81) *as mentioned above*
(82) *as you'll see*

11.2 Homework

Please, watch the student presentation from Monday of Blackboard. I've also finally uploaded the other presentations I have received.

No other homework this week. Keep reading, keep thinking about topics for your term paper. At some point during your studies, you should make 'homework' your own responsibility. :)

11.2.1 Tip of the day

One of the more annoying steps during the data acquisition via our server is getting the concordance or frequency lists to your computer. In the lecture, one rather convenient method was introduced, but you can do even better. You can set up a secure connection via WebDAV, and integrate your server files into your local file browser. If you follow the steps illustrated in the tutorials below, you will have your server space show up in your files as though it was a drive on your computer. This can save you a lot of time and hundreds of clicks.

- Setup Windows: click here
- Setup Linux (Ubuntu): click here
- Steup MacOS: click here

For general information, click here

BONUS Tip:

You can use your server space to create your own website. So if you've played with the thought of setting up your website, this is a convenient way to experiment. Normally, you have to acquire your own domain and server space, but here you can get right to it. Bear in mind that the website you set up there is gonna expire with your uni account. It can be a nice playground, though, for a real future website. Or maybe you want to learn some HTML/CSS or even PHP or Javascript.

12 Constructions

12.1 Construction Grammar

12.1.1 Construction

- Constructions are a phenomenon between purely morpho-syntactical and purely lexical patterns
- lexico-grammatical
- also form-meaning pairings

12.1.2 Examples

- Ditransitive construction

(83) She gave him a happy smile. (COCA)
(84) Mail me the results.

- *into* causative (cf. Stefanowitsch & Gries 2003)
  - [VERB] [SOMEONE] into [VERB*ing*] [something]

(85) … to **pressure the government into accepting** …
(86) … having been **tricked into thinking** Shatov is a danger (BNC)
(87) There are few things worse than **being bludgeoned into reading** a book you hate. (BNC)

12.1.3 More than the sum of its parts

Discuss: What are the crucial meaning components of the parts in bold.

(88) I looked at the drummer, and he **dropped me a sick beat**.

In 88, there is an abstract sense of "transfer" of "offer" which is not part of the semantics of its component parts. The beat is now "mine" to dance or jam to. *drop* in this sense is used

roughly for “suddenly starting music”, *me* is a deictic pronoun simply pointing to the speaker, and a sick beat stands for “an enjoyable, highly danceable/jammable” rhythm. The meaning of transfer can only come from the construction itself. Even if we constructed a clause with nonsense words, it is likely to feel like a transfer.

(89) He gnubgnarfed us a tiwilbemba.

It works as a ditransitive because we associate the meaning of transfer with the construction itself.

Now, consider the following example of the *into* causative

(90) In the end, they **embarrassed me into deleting** the photos

embarrass again does not necessarily cause a specific action. It is most often used to describe an internal state of mind. Likewise, *into* as a directional preposition does not have a meaning component of causation. We could assume a “causative” *into* as a sense or homonym of into. However, the issue would be that we would inflate our model of the lexicon with a new sense for every use we encounter that is not explained by the meaning components. A construction grammatical model would solve this more elegantly again. The causative meaning comes from and is learned from the construction itself. Again, we can play with those associative patterns creatively:

(91) He gnubgnarfed us into reading a tiwilbemba.

12.2 Organizational matters

12.2.1 Poster session

See appendix for general information on academic posters.

Format:

- Posters and Video contribution will be available in a repository and subsequently uploaded to Blackboard
- Presentation groups[12] will be split up into 5 concurrent sessions based on topics
- Every group has 5 minutes for a poster pitch with 5–10 more minutes for discussion
- In the main lobby there will be a schedule with all presentations
- Viewers (and also presenters when it’s not their turn) can switch between sessions freely

For presenters:

- Submit you contribution by **Saturday 24:00**
- if you made a poster submit it as a **.pdf** file or in an **image format** (.bmp, .jpg, .svg,...), **not** .pptx or similar

12.2.2 Term paper

- Inform me about your topic. This will function as ‘registration’
- Submissions before 21.08.20 will receive quick grading and a full feedback on request. After this date, I have limited capacity.
- I will not accept submissions after **05.10.2020** unless there has been some clear communication about it via email
- Failing registration or submission will not cause a fail, but require you to retake the course

[12] If you are alone, you are a one-person group. ;)

12.3 Homework

Work on your posters or term papers.

Otherwise, no homework, just a little idea I thought would be cool. :)

Send me a selfie of yourself in your 'home office'. I would like to compile a little class photo and upload it so everyone has a memory of this weird semester, and at least has an idea of who was behind those names. If you don't want to send a selfie, send me some other photo or even any other avatar to represent yourself. It also doesn't need to be at your desk. Surprise me with your creativity.

12.3.1 Tip of the day

Today, just some reflections on a general mindset I think people can profit from:

No matter your skill level: **re-read and re-watch basics over and over.**

Instructors have a different perspective and very often explain aspects that seem important at their own skill level. Sometimes there are realizations of the type: "I should have known that when I started", or, "now that I know x, y becomes so much clearer". Very often, however, this is a fallacy, and that type of information is not yet useful to a beginner at all. Therefore, most introductory materials have a lot to offer to advanced learners as they offer insight into the thinking of a fellow-learner. I, personally, still go over introductory materials again and again, be it in linguistics, statistics, programming or whatever I need in my day-to-day job. People who stop with that, I believe, lose track of what's important really quickly. They also might not even be aware that they don't have sufficient understanding of some of the 'basics' in their field.

So re-read, re-watch, re-visit. If you think, you know your way around in your field of interest, go back and reflect on it. There will be aspects you have overlooked. And, if you feel like your still a novice, it'll help anyway. Worst case: you have the same joy of discovering the facts and feelings that lead you to your field in the first place. It's never a waste of time. :)

Which leads me to the practical conclusion: if you struggle to find something in linguistics that is worth writing about, return to the beginnings, skim through introductory videos, textbooks, slides, etc. If you're not yet brimming with ideas and vibrating with an urge to find out more about language, go back to the basics. Maybe you discover things you didn't see when there was an exam in your neck.

13 Term paper guide

In the following sections, I am going to provide a short overview on what you should consider when writing your term paper. In general, your paper is a miniature research paper, and any of the course readings can serve as a model.

Here are some basic guidelines:

- Inform me about your topic. This will function as 'registration'
- Submissions before 21.08.20 will receive quick grading and a full feedback on request. After this date, I have limited capacity.
- I will not accept submissions after **05.10.2020** unless there has been some clear communication about it via email
- Failing registration or submission will not cause a fail, but require you to retake the course

13.1 How to hand in

Via email, in .pdf format.

For archiving purposes, we also need a hardcopy with a signed declaration of integrity (see below). With Corona going on, this is not urgent. Just drop it in our post box (outside JK29/245) whenever you are back on campus. You can also send it via post. Just staple or clip the pages, nothing fancy, no binder/folder/plastic wrap.

13.2 Requirements

In your term paper, your task is to develop an interesting research question, find literature about a linguistic phenomenon, and extract data that you then analyze and interpret.

1. **Form:** A good paper adheres to general conventions for writing papers (see below), and also linguistic conventions (cf. tip of the day #3).
2. **Language:** A good paper is written in an academic style. The more academic language you have read, the easier this will be for you to emulate. Of course, you should also follow proper spelling and punctuation conventions. Use clear and concise language and build up your arguments logically and easy to follow.
3. **Terminology:** Naturally, you should use linguistic terminology correctly, i.e. in accordance with convention. One of the most common mistakes, however, is not identifying the right places to use terminology, which is often a sign of bad literature research or a lack of linguistic knowledge. If a structure has a name in linguistics, use it. For example, an adverb referring to time is a temporal adverb; an adjective appearing in front of a noun is an attributive adjective, etc…
4. **Operationalization:** You need be able to make the linguistic concepts you discuss measurable. In most cases, this comes down to the question of, "how can I count occurrences of x". If you use counts, you need to make sure these counts represent your phenomenon. If you code data, you need to take decisions that are conceptually motivated.
5. **Methodology:** Your paper should make use of the empirical methods we have learned over the course of this semester. A good paper not only gathers valid corpus data reproducibly, but also describes them with the right metrics. An excellent paper is also aware of statistical significance.
6. **Line of argument:** A good paper builds up a compelling line of argument that is aware of limitations, without sacrificing the meaningfulness of the study. Common mistakes are on both extreme ends of a scale: either completely refuting the validity of the applied method or data; or over-generalizing results and accepting a hypothesis without sufficient evidence.

13.3 Form

13.3.1 Requirements

- **Length:** c. 2000 words;
- **Language:** English

You can write about a phenomenon in a language other than English, but the language of the paper should be English.

2000 words is a rough guideline. There are no automatic penalties for staying below or exceeding this limit. Papers that are shorter usually suffer from either a lack of literature discussion or a lack of data. Papers that are a lot longer usually fail to narrow down the topic enough, ending up too ambitious.

13.3.2 Typography

Stay consistent! That is almost the only rule. Below are some conventions you should stick to.

Page formatting:

- Separate title page
  - Includes the title of your paper, your name, Matrikelnummer, course ID, instructor, semester, and date.
- Separate table of contents
- Separate bibliography
- Page numbers start on page 1 of the Introduction

Text formatting:

- Reference to words and phrases in text: *italics*
- emphasis in examples: **bold**
- emphasis in direct quotes: underlined
- examples consecutively numbered (unique number for every example)
- tables and figures should be numbered[13]

Citation and bibliography style: Please use the following style sheet.

- Unified Style Sheet for Linguistics
- .csl file for use with Latex or Markdown

13.3.3 Structure

1. Introduction

Contains your research question, introduces the main terminology and provides an overview of your paper. Almost all important information should already show up here, including the most important results.

2. Main part

Your are free to create any number of subsections you think are necessary. A good rule of thumb: 3-3-3. Have 3 main sections each containing 3 paragraph with each three relevant arguments. In 2.1 you typically define and discuss terminology and concepts with the help of literature references, 2.2 is for explaining your methodology and 2.3 is for the analysis.

3. Conclusion

The little brother of the introduction. Should sum up everything, argue whether the research question was answered, hypotheses supported or rejected; and consider drawbacks of your method and potential for further study.

13.3.4 Appendix

It is good practice to append queries, and scripts you have used. For longer analyses, researchers might even create a repository on Gitlab, Github or Bitbucket with all the files in it. In your case, this either does not apply or is probably overkill. Your only concern should be: is my data analysis reproducible given my explanation? If you want to attach your queries or whole data sets, put it in an appendix. If it exceeds 3-5 pages, put it in a file and email or upload it.

[13] Given the short length of the term paper, you don't need a list of figures and tables

13.3.5 Declaration

Finally, some bureaucracy. As a last section, you have to add a declaration of academic integrity, in which you testify that you did not plagiarize anything and that you have not handed in the same paper anywhere else. Following is an example (German version since it is German bureaucracy).

```
Erklärung

Name:
Adresse:

Hiermit versichere ich, dass ich die vorliegende Hausarbeit selbstständig
verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt
habe; alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen
wurden, kenntlich gemacht sind und die Arbeit in gleicher oder ähnlicher Fassung
noch nicht Bestandteil einer Studien- oder Prüfungsleistung war. Unterschrift
der Verfasserin / des Verfassers: Datum:
```

Appendix

The following sections contain some miscellaneous information in various topics around the course. Though, not strictly necessary to finish the course, some of the articles might be interesting for some of you.

Academic posters

Your poster presentation (information on the presentation day) is essentially a progress report of the project you are working on. Its structure is mostly identical to that of an academic paper. There should be an introduction (usually at the top), and a conclusion (usually at the bottom). You should also reference research literature and include a bibliography.

The main difference is that, if possible, there should be less text and more examples, figures, and tables. Ideally, you have already explored some corpus data and have some preliminary results. Here is a list of elements that could be at the heart of your poster.

Linguistic data:

- Numbered examples that illustrate your phenomenon (ideally from your data set)
- Concordances
- Frequency list
- Tables with counts for individual categories

Visualizations

- Bar charts
- Stacked bar charts
- Pie charts
- Scatter plot …

Conceptual Figures

- Flow charts
- Venn diagrams
- Models

Layout

When it comes to the design of your poster, it is mostly up to your creativity. Posters are usually A0, so quite large. For layouts, just google academic posters or linguistics posters. In academia, your institute or university usually has a corporate design and might even provide templates. Corporate designs include logos, colors, fonts and other instructions of varying specificity (e.g. FU corporate design).

More common layouts include headers and footers. The header includes the title, logo, names of the authors, their affiliations, and contact information. The footer includes references, acknowledgements, footnotes. This provides a frame for the main body, that has numbered sections, just like a paper. Sometimes people include an abstracts at the beginning that is a summary of the project.

I have provided a very simple template on Blackboard for you. Feel free to use it.

Programs

Most commonly, people create their posters in presentation software like Powerpoint / Impress. If you are already familiar with image editing programs like Photoshop or programs for graphical design, these might be an option for you. The most powerful, and extensible options are Latex or Markdown, which offer great functionality when it comes to references, bibliographies, cross-references, numberings and captions. For a beginner, it might be extremely difficult to work with those tools without mouse drag-and-drop, but you can just download example files from places like Overleaf (e.g. here) and just throw in your contents.[14]

Starting from scratch might be daunting. However, there are countless templates online. Just search for one created with your tool of choice, pick one you like, and modify it if necessary.

Command line tricks

Working through the command line, i.e. working with a text-only interface from a terminal, is a very powerful way to do precise and flexible data manipulation. Unfortunately, the command line has fallen out of use for a wide variety of reasons, non of which having to do with it being in any way inferior or more difficult. As a matter of fact, it is easier to do certain operations, in particular producing automated and reproducible processes, and it is the preferred method for a large group of scientists (not only programmers) and even just regular “power users”.

In the following section, I will gather some tricks that make it easier for you to work with command line tools like CQP. A terminal only handles text in- and output so mouse functionality is limited. Some keys and key combinations also migh show unexpected behavior.

13.3.6 The shell

The shell is what is running in your terminal and interprets your commands. It is the counterpart of the background process of Windows, Linux or macOS that draws windows.

13.3.7 Copy & Paste

Depending on your terminal program you might want to try the following things.

- `ctrl + shift + c` (copy) and `ctrl + shift + v` (paste)

[14] I do not recommend making posters or slide presentations in Latex as an absolute beginner. With that said, you can learn a great deal about how styling works if you try. I forced myself to prepare a whole seminar in Latex beamer, and it was a great learning experience. But that was when I was already highly dedicated to switching away from Word and Powerpoint. It can be frustrating and requires patience.

- middle mouse button (in Linux and macOS)
- Windows only: right click, (both copy and paste)

`ctrl + c` and `ctrl + v` have different meanings in most terminals. `ctrl + c` e.g. cancels the current process.

13.3.8 Colors

It might sound weird, but the default color scheme in a terminal is probably one of the main reasons people find working in the command line scary. Black on white or white on black (blue for Powershell) are ugly, hurt your eyes, and can make focussing for longer time periods difficult. My suggestion is, therefore, to switch this to a low contrast color scheme. Light gray text on a darker gray background is what I personally can work with best.

- Windows Powershell: right click on the top bar → properties → colors
- Windows Terminal: See here
- Putty: color menu is right on the left on the start screen
- macOS: right click on "Terminal" in the panel → properties →pr

There is a whole parallel universe of people for whom terminal color schemes are an art form (check out this subreddit). Since CQP does not produce colored output (yet), it is unfortunately not too useful for us.

13.3.9 Eastereggs

- Star Wars in the terminal: `telnet towel.blinkenlights.nl`
- Dancing parrot: `curl parrot.live`

Why discord?

You might wonder why I have opted to use Discord rather than some of the other alternatives, especially since it is rather a platform designed for gamers. I have tested a variety of software, including Webex, Zoom, Skype, Jitsi, Tox, etc… There are several reasons why I have decided in favor of Discord.

1. Integration Discord has everything combined in one place. We have chat rooms, presentations via live stream/screen cast, and voice+video chat capabilities. You can also interact with your peers without figuring out anyone's user name or mail address. Having everything in one place should minimize the amount of things that can go wrong. And there **will** be bugs and issues, and I don't want to become a professional tech-support. It should also be less confusing for students, especially since you have other courses that use other software.

For this course, you just have to remember: Discord for interaction, website for materials. I can even offer my office hours in the same place. ## Why discord? {-#why}

You might wonder why I have opted to use Discord rather than some of the other alternatives, especially since it is rather a platform designed for gamers. I have tested a variety of software, including Webex, Zoom, Skype, Jitsi, Tox, etc… There are several reasons why I have decided in favor of Discord.

1. Integration Discord has everything combined in one place. We have chat rooms, presentations via live stream/screen cast, and voice+video chat capabilities. You can also interact with your peers without figuring out anyone's user name or mail address. Having everything in one place should minimize the amount of things that can go wrong. And there **will** be bugs

and issues, and I don't want to become a professional tech-support. It should also be less confusing for students, especially since you have other courses that use other software.

For this course, you just have to remember: Discord for interaction, website for materials. I can even offer my office hours in the same place. Trust me, I am all for software minimalism, but priorities change when coordinating large groups.

2. Persistence On our server, I think we can achieve the closest to a coherent class room feeling. Most of the alternatives are based around meetings. The host sets up a meeting and people have to join or get invited every time. This can be automatized to varying degrees, but it still means some setup overhead every single time, which is prone to error. On our Discord server, you just have to setup once, after which you just start the application at seminar time or whenever you have a question. The text chats are persistent and can function like a forum or blog.

3. Sensible defaults I believe that a good online course should keep the distractions at a minimum. There are enough distractions at home. And in my opinion, video is a big distraction during a presentation. Not only do the videos of other participants potentially distract from the actual presentation, but your own video itself does. Everyone knows why their opposite during a video chat is always looking at one of the edges of their screen ☺. We're all guilty of that.

Some of you might have been to larger video conferences on platforms like Webex and Zoom. If so, you know how distracting and annoying some of the default features can be. Like shifting video focus to active speakers, which happens when someone coughs, moves their chair, bumps into their microphone etc. Eliminate video, problem solved. I do think that seeing each other improves the experience and makes everything more personal. But for this, I prefer smaller group-based video calls in addition to regular presentations.

4. Limitations of alternatives Other than missing features, there are some major limitations that rendered the alternatives useless for my purposes.

References

Altenberg, Bengt & Sylviane Granger. 2001. The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics*. Oxford University Press 22(2). 173–195.

Berg, Thomas. 2000. The position of adjectives on the noun–verb continuum. *English Language & Linguistics*. Cambridge University Press 4(2). 269–293.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*. Oxford University Press 25(3). 371–405.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics*. Mouton de Gruyter 37(4). 575–596.

Davies, Mark. 2008. *The corpus of contemporary American English: 450 million words, 1990-2012*. http://corpus.byu.edu/coca.

Deignan, Alice. 2005. A corpus linguistic perspective on the relationship between metonymy and metaphor. *Style* 39(1). 72–91.

Deignan, Alice. 2006. The grammar of linguistic metaphors. In Anatol Stefanowitsch & Stefan Th. Gries (eds.), *Corpus-based approaches to metaphor and metonymy*, 106–122. Mouton de Gruyter.

Divjak, Dagmar & Stefan Th Gries. 2009. Corpus-based cognitive semantics: A contrastive study of phasal verbs in english and russian. *Studies in cognitive corpus linguistics* 273–296.

Geeraerts, Dirk. 2015. Lexical semantics. In, *International encyclopedia of the social & behavioral sciences*, 931–937. Elsevier Ltd.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. 1st edition. University Of Chicago Press.

Gries, Stefan. 2009. What is corpus linguistics? *Language and Linguistics Compass* 3. 1225–1241. doi:10.1111/j.1749-818X.2009.00149.x.

Horsmann, Tobias, Nicolai Erbs & Torsten Zesch. 2015. Fast or accurate? – a comparative evaluation of pos tagging models. In, *Proceeding of the second italian conference on computational linguistics*, 166–17. Trento, Italy: Accademia University Press.

Justeson, John & Slava M Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational linguistics*. MIT Press 17(1). 1–19.

Kennedy, Graeme. 1991. Between and through: The company they keep and the functions they serve. In Aijmer K. & B. Altenberg (eds.), *English corpus linguistics*, 95–110. London: Longman.

Kennedy, Graeme. 2003. Amplifier collocations in the British National Corpus: Implications for English language teaching. *Tesol Quarterly*. Wiley Online Library 37(3). 467–487.

Maddieson, Ian. 2013. Front rounded vowels. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/11.

OED Online. 2020. http://www.oed.com/; Oxford University Press.

Ross, John R. 1972. The category squish: Endstation Hauptwort. In, *Papers from the eighth regional meeting of the chicago linguistic society*, vol. 8, 316–328. Chicago Linguistic Society.

Schmid, Hans-Jörg. 1996. Introspection and computer corpora: The meaning and complementation of start and begin. In, *Lexicographica, symposium on lexicography vii*.

Schmid, Helmut. 2013. Probabilistic part-of-speech tagging using decision trees. In, *New methods in language processing*, 154.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Draft.

Stefanowitsch, Anatol & Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*. John Benjamins 8(2). 209–243.

Stefanowitsch, Anatol & Stephan Gries. 2009. Corpora and grammar. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: An international handbook*, 933–951. De Gruyter Berlin, Germany.

The British National Corpus, version 3 (BNC XML Edition). 2007. http://www.natcorp.ox.ac.uk/; Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Weisser, Martin. 2016. *Practical corpus linguistics: An introduction to corpus-based language analysis*. New York: John Wiley & Sons.