Levels of Linguistic Analysis II: Morphology

Alexander Rauhut, M.A.
AG Struktur des heutigen Englisch
Freie Universität Berlin

Winter Semester 2020

Contents

1 Welcome **2**
Syllabus . . . 2
Contact . . . 3
Links . . . 3
Course Requirements . . . 3
Weekly workflow . . . 4

2 Form and meaning **5**
2.1 Linguistic questions . . . 5
2.2 Questions in morphology . . . 5
2.3 Course Aims . . . 6
2.4 Homework . . . 7

3 Word Classes **8**
3.1 Parts of speech . . . 8
3.2 Types and Tokens . . . 9
3.3 Tip of the day . . . 13

4 The Lexicon **13**
4.1 Lexemes and lexical fields . . . 14
4.2 Frequency and memory . . . 15
4.3 Homework . . . 16

5 Collocation **17**
5.1 Tip of the day . . . 18

6 Metaphor **19**
6.1 Metaphor and quantitative evidence . . . 19
6.2 Metaphor and Cognition . . . 21
6.3 Homework . . . 22

7 Cognition and Categorization **23**
7.1 Models . . . 23
7.2 Examples . . . 24
7.3 Tip of the day . . . 25

8 Irregular inflection **25**
8.1 Homework 25

9 Metonymy **27**
9.1 Comparing Apples to Mangos 29
9.2 Tip of the day 30

10 Productivity **30**
10.1 Tiwilbemba 34

Appendix **36**
Term paper 36
Academic posters 39
Command line 40

References **41**

1 Welcome

Welcome to *Levels of Linguistic Analysis II: Morphology*! 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.

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.

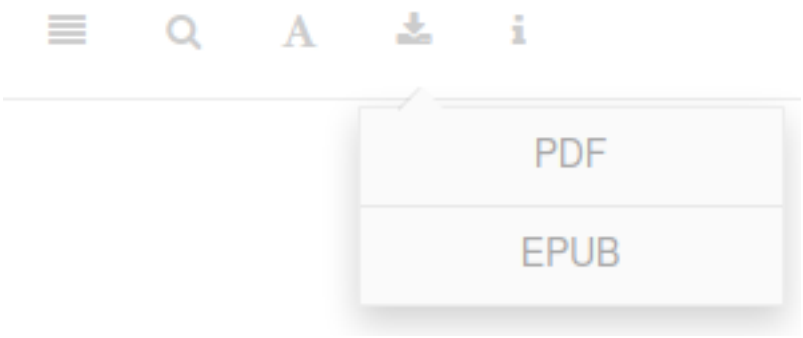


Figure 1: Just click on the download symbol in the top left corner.

Syllabus

	Date	Topic	Main Reading	Homework[2]
1	04.11.	Welcome		
2	11.11.	Form and meaning	Stefanowitsch (To appear) ch. 1-2	#2
3	18.11.	Word classes	Stefanowitsch (To appear) ch. 3	

[1] Experimental, not perfect but sort of works

	Date	Topic	Main Reading	Homework[2]
4	25.11.	Lexical Relations	Justeson & Katz (1991)	
5	02.12.	Collocation	Kennedy (2003)	
6	09.12.	Metaphor and metonymy	Deignan (2005)	#6
7	16.12.	Focus I: Derivation	Kaunisto (1999)	#7
8	06.01.	Focus II: Irregular inflection	Anderwald (2011)	#8
9	13.01.	Focus III: Conversion	Deignan (2006)	#9
10	20.01.	Focus IV: Productivity	Plag, Dalton-Puffer & Baayen (1999)	#10
11	27.01.	Focus V: Morphosyntax	Rosenbach (2003)	#11
12	03.02.	Project Day 1	t.b.a.	
13	10.02.	Project Day 2	t.b.a.	
15	24.02.	Final Discussion		

Contact

- 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

... to be continued

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**.

[2] If link leads nowhere, the assignment is probably not ready yet. If it's late, shoot me an email.

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

Weekly workflow

There are three basic components to our seminar:

1. This website
    - Main resource
    - Homework
2. Live sessions
    - Weekly presentations
    - Student presentations
        - Student presentations
3. Self Study
    - Reading assignments
    - Practice

I. This Website This website is going to be the main hub for information. It will essentially replace most PDF materials you are familiar with from regular semesters, such as presentation slides. You will find all course information, the syllabus, bibliography, and tips and tricks, which you can easily navigate with the sidebar.

All **homework** will be published here. I'd recommend you bookmark the syllabus, because everything that is relevant weekly is linked from there.

The main sections will replace presentation slides. They will be written in the form of short articles that pick up some major points that came up during the live session. They also might go a bit deeper into certain subjects.

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. Inform me about broken or misdirected links. :)

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 or found online
- 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.

II. Live streams Every week at the scheduled seminar time—Mondays 16:00-18:00—, I will live-stream my main presentation. 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. Self-study The bulk of the work, you have to do by yourself. This is nothing special about an online semester. If you look up what an ECTS credit represents you will find that it is work load measured in time. If you then subtract the little time we use during live sessions, you'll realize how much time is left for you to prepare for every week, study the readings, discover your own further readings, do research for your own project, or practice. Ideally you should have read even more than just the recommended literature by the end of the seminar. Reality check if you find yourself skipping entire readings or procrastinating homework assignments.

2 Form and meaning

Main goals of week one is to get comfy with the online format and get to know each other. We also want to refresh our memory about linguistics and get an overview about what is coming in the following weeks. I asked you why you picked morphology and what you expect from the class. We also discussed what questions are driving research in morphology.

2.1 Linguistic questions

Throughout the course, wee are going to discuss various topics mostly from—but not restricted to—the field of morphology, 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?

2.2 Questions in morphology

A lot of focus in morphology is on the relationship between form and meaning. The further we go to the grammar side of linguistics, the harder it becomes to use an intuitive concept of meaning as a starting point. It seems rather easy to determine the meaning of the word *cat*. A simple definition usually satisfies. We can characterize *cat* as a word. It is phonologically and morphologically distinct from other words. It is also a morpheme. It can serve as a root for more complex words, such as *cats*. The morpheme is, "the smallest meaningful unit," remember? So, what about the meaning of *-s* in *cats*? If the root morpheme has a meaning, the plural suffix has to has one as well. A straight forward answer would be: plural, more than one. You were introduced to this as "grammatical meaning." Lexical words have lexical meaning, function words and inflectional affixes have grammatical meaning.

Matters still seem simple enough until we enter more complicated territory in English. Consider the *-ed* suffix like in the verb *helped*. Most people would be quick to call that the "past tense" and *-ed* is the past tense ending. What is complicating things now is the fact that *-ed* is also used in present tenses, or passives.

(1) Present perfect: I have watched the show already.
(2) Past perfect: I'd never laughed so hard in my life before that.
(3) Passive: I'm being observed right at this moment.

First, we could assume that there are, in fact, multiple *-ed* endings. In irregular verbs, there is another form, the past participle: *write*, *wrote*, *written*. It would only make sense to assume that there is a *-ed* past form and a *-ed* past participle form. That means we have a kind of homonymy here. One form that has two different meanings. So what is this meaning? Is it "past?" Does the event denoted by the verb happen in the past relative to the time of utterance?

Consider these examples:

(4) Backshifting: Trump has claimed there **was** evidence for fraud.
(5) Unreal Conditional: If I **met** him tomorrow, I would slap him with a fish.
(6) Optative: I wish people **believed** in science.

In neither of these examples does the event occur in the past. Therefore, we have to assume a more abstract function of the past tense suffix *-ed* that allows for all these use cases. More importantly, the surrounding forms, auxiliary verbs (passive *be*, perfect *have*), conjunctions (conditional *if*), and even certain classes of verbs (reporting verbs, such as *claim*). In fact, a linguistic **form** is not restricted to a word or morpheme, but may also include larger structures or even abstract schematic structures. For example, the grammatical function "present perfect" is fulfilled by a wealth of different forms, with optional slots, different possible word orders, etc.

The popular opinion in linguistic theory, especially in Usage-Based Linguistics, has shifted more and more towards seeing meaning and function as two sides of the same coin. Function determines meaning and meaning is in the end a function.

Here are some broader questions that we will encounter during this class:

1. Why do we have multiple functions that seem to be encoded in the same form? (homonymy)
2. Why do we have multiple forms that seem to do the same thing? (synonymy)
3. Is a morpheme even a part of our cognitive reality or can we find better units of description? Are there better models for certain areas of grammar?

2.3 Course Aims

2.3.1 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**

2.3.2 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

2.3.3 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

2.4 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. Setup a VPN connection to the university network.
   - Setup guide
2. Download the main readings online and download them.
   - Google Scholar (you can search for authors by typing author:*name*)
   - Primo
   - Sometimes papers or entire books are uploaded on the authors own website, so regular search engines help sometimes
3. Make a note for every reading you couldn't find. Some require a bit of digging, but they are all out there.

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.

2.4.1 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

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

3 Word Classes

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.

3.1 Parts of speech

3.1.1 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.

3.1.2 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)

3.2 Types and Tokens

3.2.1 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:

(7) 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.

(8) 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]

(9) They wore it only because their **mothers-in-law** insisted. (BNC[5])
(10) I always thought it was **mother-in-laws** that cause the problem. (COCA[6])
(11) 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:

[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)

(12) *going to*
(13) *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**

3.2.2 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.

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
```

3.2.3 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)

3.3 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 since not everything is digitalized.

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. Platforms like Primo and Google scholar provide bibliography entries in various styles and formats. 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.

4 The Lexicon

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.

4.1 Lexemes and lexical fields

4.1.1 Lemma

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

(14) be, am, are, is, were, was, been, 's, 'm, 're, ?being
(15) cut, cuts, (cut, cut), ?cutting
(16) tree, trees, tree's, trees'
(17) nice, nicer, nicest
(18) 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.

4.1.2 Distribution

Information about the frequency of a word or its forms can already be very informative. We can extend this idea easily and look at the larger units a word or lemma occurs in. Frequency information about the **distribution** is much more complex, but is based on the same underlying concepts and measured with the same tools.

We have seen already from Justeson & Katz (1991) we can see that the distribution of adjective pairs plays a crucial role in the formation of antonym pairs. There, the deciding factor was whether they occur together in the same context. Different form same context. We could flip this around and look at words with same form that occur in wildly different contexts. A special case of this is **homonymy**.

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?

(19) Maybe I'm a fruitarian **bat**
(20) … with a straighter **bat** than some of the Englishmen
(21) The unfortunate starved **bat** was then returned
(22) 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.

4.1.3 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**.

(23) law and …?

- order

(24) good or …?

- bad, evil

(25) the number of the …?

- ??beast

(26) 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.

4.2 Frequency and memory

4.2.1 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.

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

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.

4.2.2 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)
- *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.

4.3 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.

4.3.1 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!

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.[8] 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.

5 Collocation

Coming soon: I will extend this section soon with some hands-on examples.

Major Concepts

- Collocation and Co-occurrence
- Frequency and relative frequency
- Reproducibility and coding

Perception of Quantities

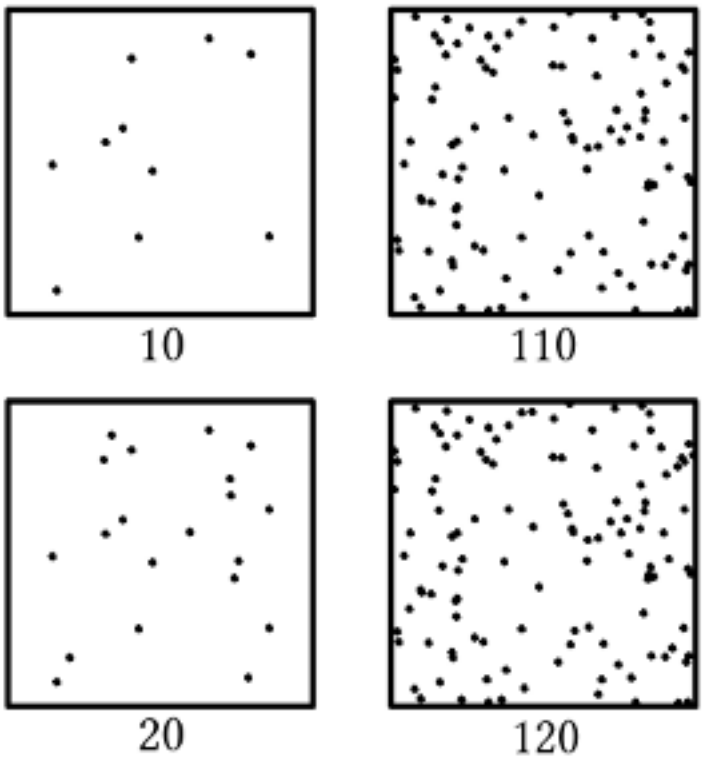
Frequency of occurrence is used as an **indicator**

How do we perceive quantity?

Weber-Fechner Law

- human perception is based on ratios, not absolute values
- absolute differences become exponentially less informative

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



5.1 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:

(27) *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)

(28) *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 Frequency and scales

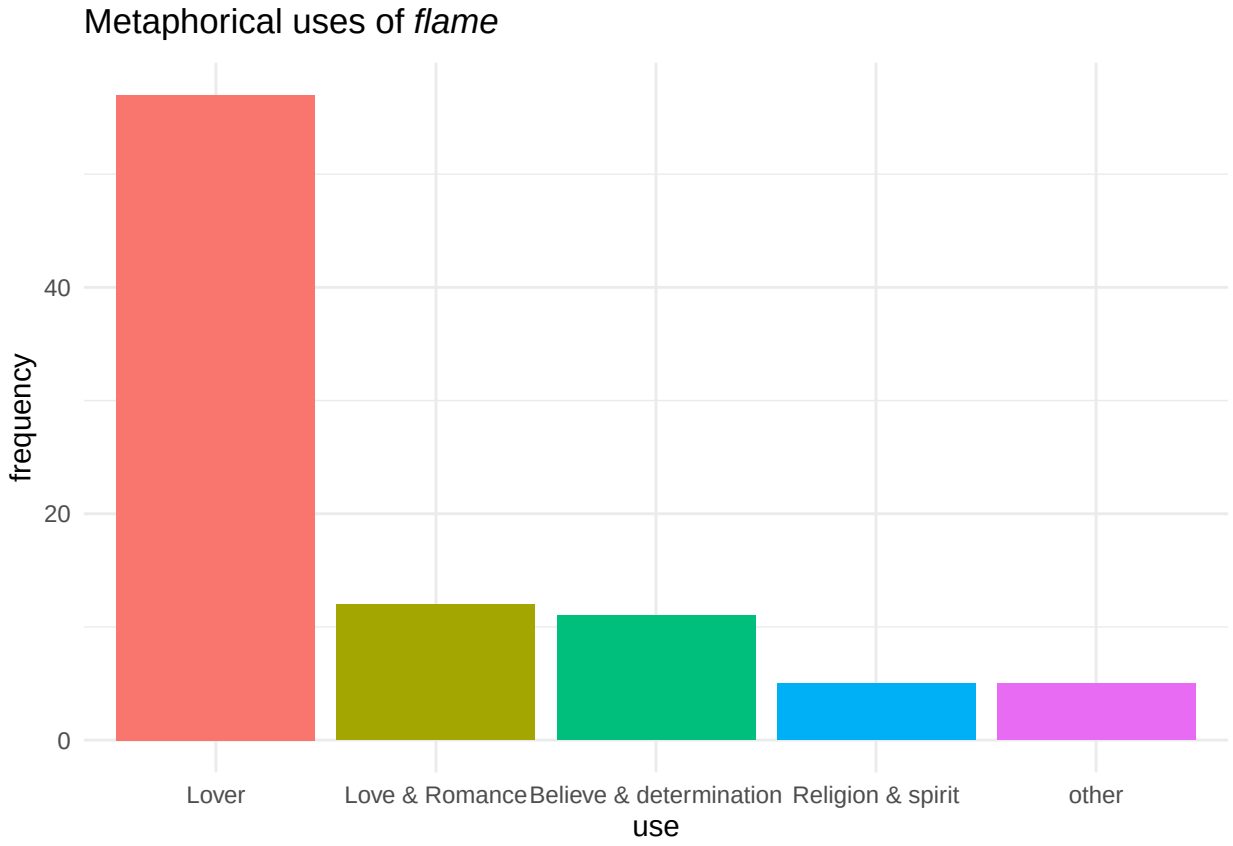
- Absolute frequency
  - Basic measure
  - Should always be reported since everything else is based on it
  - Sometimes hard to visualize
  - Hard to interpret across different sample or category sizes
- Relative frequency
  - Absolute frequency divided by all occurrences
  - Either between 0 and 1 or 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%

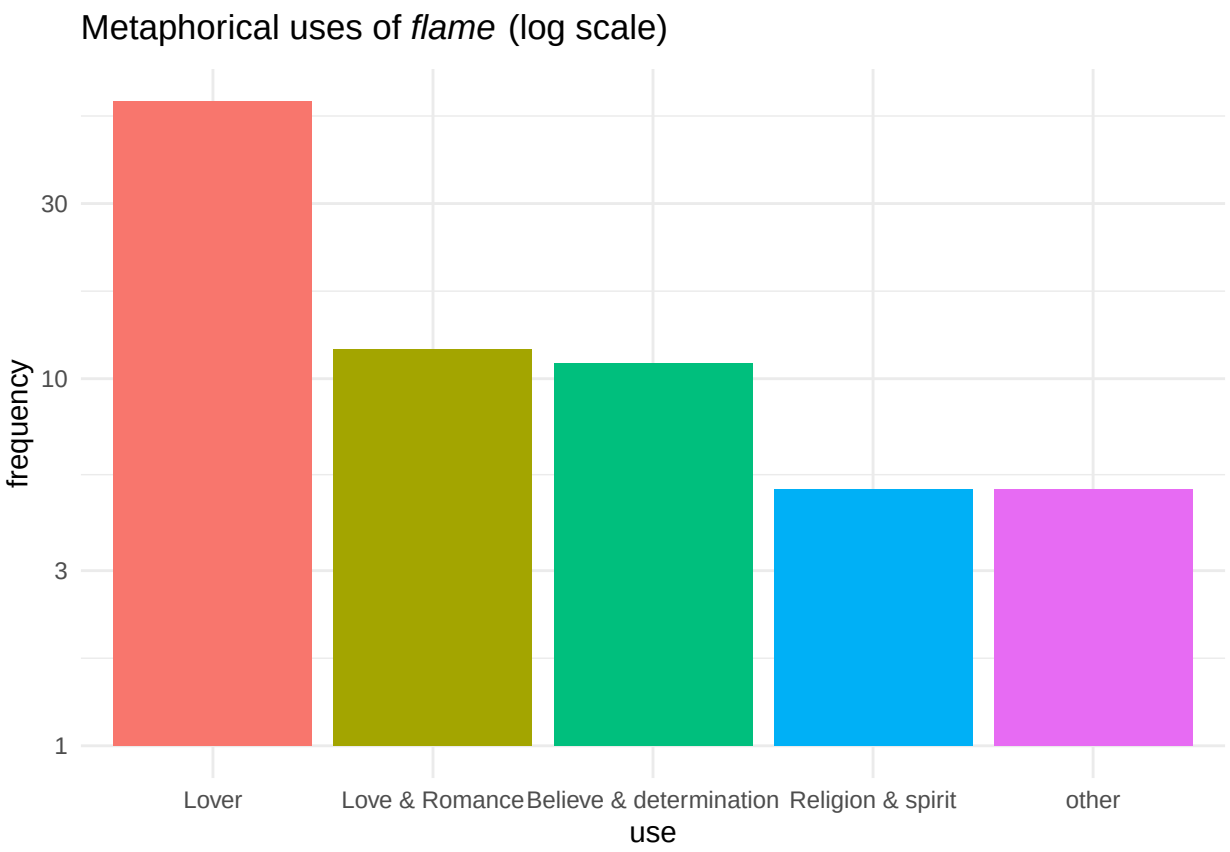
6.1.3 Frequency and scales

- 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

6.1.4 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

6.3.1 Task

Here is another paper about metaphors in corpora (Deignan 2006). Skim through it and 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. Find Deignan (2006) online and pick some tables 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 by Friday.
4. (Expert mode: Do you think the data is best presented on a log-scale?)

6.3.2 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 Cognition and Categorization

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

7.1 Models

7.1.1 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

7.1.2 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:

7.1.3 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.

7.2 Examples

7.2.1 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

```
English
----|x~~~~~~~~~ /b/
----|---x~~~~~~ /p/

Russian
x~~~|~~~~~~~~~~ /b/
----|x~~~~~~~~~ /p/

Mandarin
x~~~|~~~~~~~~~~ /b/
----|x~~~~~~~~~ /p/
----|-----x~~~~ /p /
```

7.2.2 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 [tr̥iː]
- lengthening of preceding vowels
  bet [bet]—bed [beˑd]
  beet [biːt]—bead [bːˑd]

7.2.3 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.

7.3 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.

8 Irregular inflection

8.1 Homework

During this week's seminar we started exploring irregular inflection of a) nouns and their plural forms and b) adjective and their comparative and superlative forms. For next week, I would like you to continue with this at home so that we can discuss this next week. So here are the two topics and three steps again that are involved

1. Brainstorm inflectional patterns, start to categorize them, consult background literature about common categorizations. Gather prescriptive 'rules' and descriptive regularities.

2. Do some research into the emergence of the irregular forms. Consult etymological dictionaries (e.g. the OED). Get an overview over potential tendencies.
3. Do some exploratory corpus search.

Build on your work during the seminar. The aim for next session is to develop research questions based on these overviews.

8.1.1 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 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

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

As an example of metonymy let's look at 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*.

(32) I have stuff on my Amazon as well
(33) can you get off my Amazon?
(34) 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` struc-

ture gives us a nice place to start looking into the Whatsapp-Kleenex-Tempo type of metonymy.

9.1 Comparing Apples to Mangos

As another example, let's apply the same logic as above but now for a different company name: *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"]
```

At this point, it might be worth looking into a CQP feature called wordlists: See the official tutorial for examples: Click

In an actual study, this list should not be arbitrary and best be exhaustive, meaning you should include all brands that match certain criteria you define first. A convenient and often used way to define a lower boundary to make exhaustive categories possible in the first place is defining a minimum frequency.

As a final example, let's take the above results and make 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.

9.2 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.

10 Productivity

When it comes to derivational affixes, there are 3 concepts that become very important when you work with data.

1. Analyzability
2. Semantic transparency
3. Productivity

The first aspect, is related to the question of whether a word is complex or not. There are monomorphemic (simple) and polymorphemic (complex) words based on whether you can break them up into component parts. This is straightforward in most cases:

(35) {Analyz}-{abil}-{ity}
(36) {Productiv}-{ity}

These two words are clearly composed of smaller meaningful units. Note that there is a lot of spelling variation at morpheme boundaries and there is also allomporphy. Don't confuse the two. *productive* is missing its final □e□ when it occurs with a suffix, but the phonological form changes. This is a purely orthographic convention. The change in stress from *productive* to *productivity*, however, does cause a change in the phonetic form: /pɹəd'ʌktɪv/ → /pɹədəkt'ɪv-/. Likewise, the change of the suffix {able} from /-ɪbɬ/ in *analyzable* to /-ɪ'bɪl-/ can be considered allomorphy.

Another dimension of complexity comes into play when encounter words in your data for which it isn't even clear whether they are complex or not. Some might seem complex, but require meta-linguistic knowledge, i.e. explicit, learned knowledge about the language. Common cases are loanwords like: *poltergeist*, *schizophrenia*, *statistics*, To a German speaker *poltergeist* clearly seems polymorphemic because it is a compound in German. This knowledge, however, needs to be learned explicitly by a native speaker of English, because neither component is used on its own in English. *schizophrenia* is a similar case. You can analyze the component parts academically by looking up the etymological origin. Within the English language without explicit knowledge of Latin and Greek. *statistics* is trickier. There might be a degree of analyzability here because the suffix *-ics* is common enough in English, especially for academic

disciplines. It is still different from regular derived words in that the root alone doesn't really carry a conventional meaning, at least one conventional in English. The root in statistics can be considered **semantically opaque**, which is the opposite of semantically transparent.

Now let's consider the following three examples:

(37) ceiling
(38) building (N)
(39) interesting

All three words have a very recognizable English *-ing* suffix. In that sense, they are clearly **analyzable** as polymorphemic. That means in a native speaker of English the different components might trigger multiple **associations**. They are still problematic examples if you are interested in derivations with *-ing*. *ceiling* has a similar problem to *statistics* because the root is not really **transparent** in meaning. */si:l/* does exist in a handful of other English words, such as *conceal*. So you could make the argument that it is slightly more transparent than *statistics*. This illustrates that all concepts discussed here are matters of **degree**. *building* and *interesting* are both more transparent in meaning and analyzable. You might still want to treat them separately if you look into *-ing* as derivational suffix because they are strongly **lexicalized**. That means they are unlikely to be perceived compositionally, i.e. we have them stored as a whole in our lexicon and there is little to no analogical derivation happening.

Finally, morphemes vary in terms of and how widely they are spread across the lexicon and how commonly they are used to derive new words. A highly *productive* derivational affix is used with a large variety of roots. This is true of the regular inflectional affixes in English. *-s* is used as plural suffix on most nouns and it is also the most likely candidate for a neologism. An irregular plural like *-en* in *oxen* is more like a living fossil. It is limited to certain words and never used with new vocabulary.

In derivation, productivity is not as clear cut. Let's consider the adjectival suffix *-ish* as in the examples below, taken from the Corpus of Contemporary American English (COCA) with the query `[word = ".+-ish"]`.

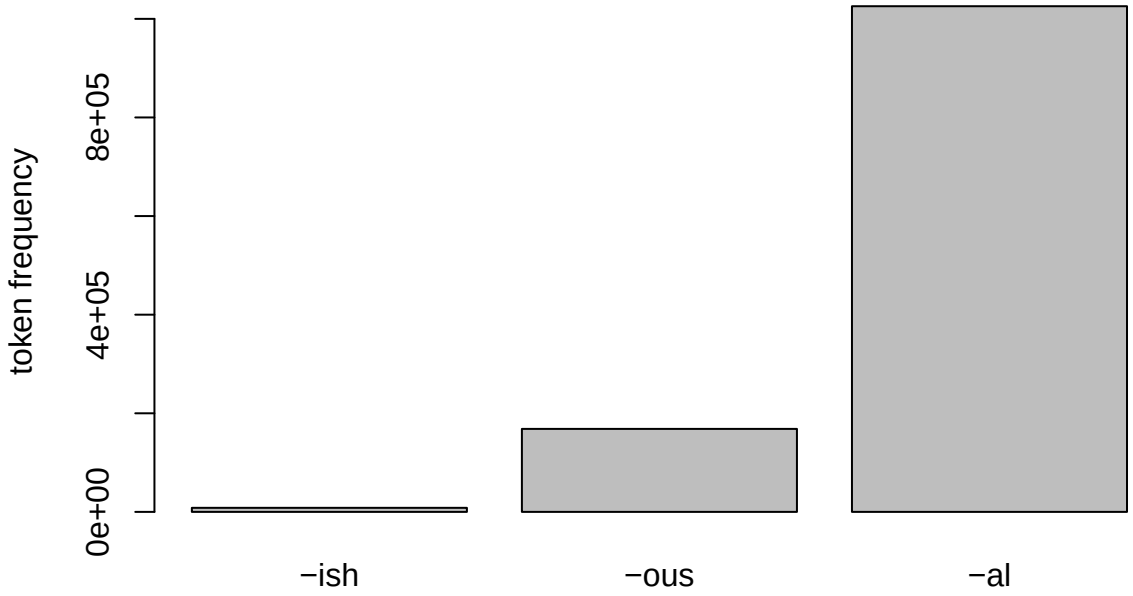
(40) I'm still feeling flu-ish.
(41) ... for a nation that was a loyal-ish member of NATO.
(42) That was selfish rather than other-ish.

In modern spoken language these types of derivations feel quite common. They may not be considered standard or formal language, which is why you will often find them hivenated. Still, intuitively, *-ish* should be considered productive.

Let's try to catch more *-ish* adjectives with a more general query. It is a good idea to exclude capital letters with `[^A-Z]` since a homonym of *-ish* frequently occurs with nationalities/language names: `[word = "[^A-Z].+ish" & class = "ADJ"]`. Let's also look at some other suffixes used to derive adjectives and compare some numbers from the BNC with some minimal cleanup:

tokens	suffix	cqp query (BNC)
8196	*-ish*	`[word = "[^A-Z].+ish" & class = "ADJ"]`
168294	*-ous*	`[word = ".{2,}ous" & class = "ADJ"]`
1025618	*-al*	`[word = ".{2,}al" & class = "ADJ" & word != "real"%c]`

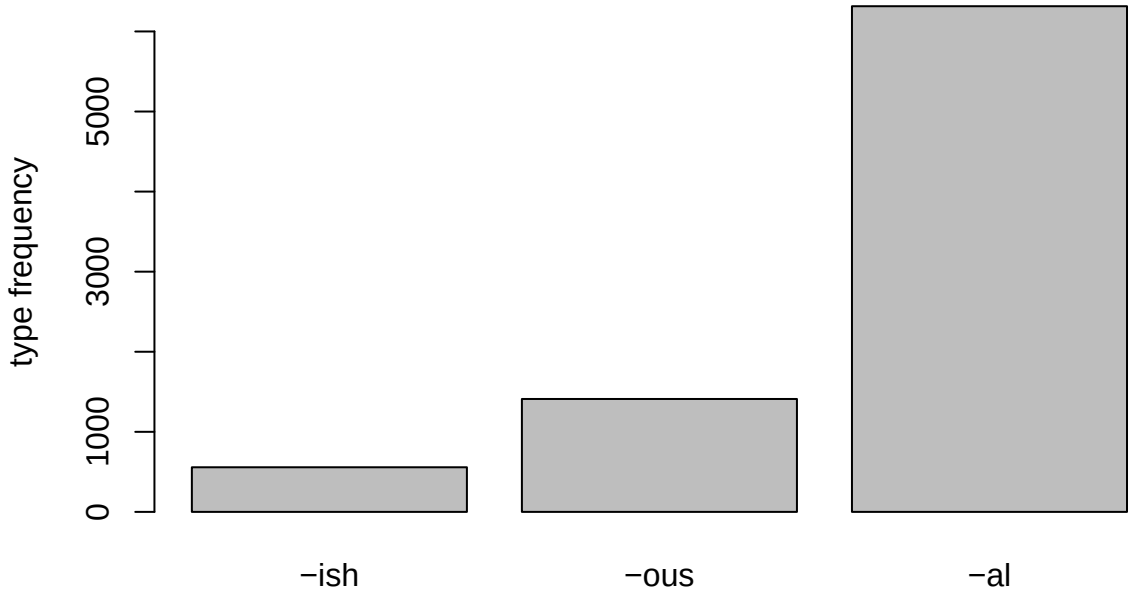
```
barplot(c(8196, 168294, 1025618),
        ylab = "token frequency",
        names.arg = c("-ish", "-ous", "-al"))
```



On a first glance, *-ish* is much less frequent in the BNC. This could have something to do with the composition of the corpus (try to restrict to match.text_mode = "spoken." Is there a difference?). Normalizing relative to sample size does not help in this case because our sample is the same across. We could also look for disproportional amounts of false positives due to our queries, but maybe this is not going to be necessary. Frequency alone is not a sufficient measure for productivity. We need to check how much of the vocabulary is covered as well. One measure that can be used is the type-token ratio. How many occurrences are there relative to how many different words they occur across? We can get the type frequency by creating a frequency list with `count` and counting how many lines it has. A quick way to do this is to use an external command line tool called `wc` with the `-l` option for line count: `count by word %c > "| wc -l"`. Note: the following graphs were made with the statistical software called R. I provided the code just in case anyone is interested.

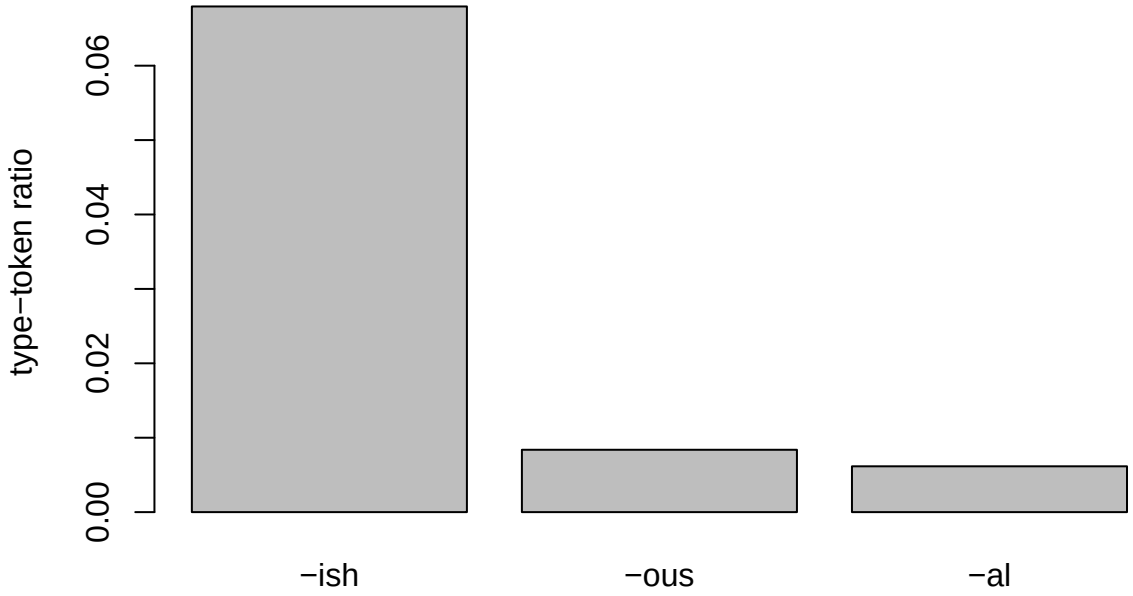
types	tokens	suffix
557	8196	*-ish*
1410	168294	*-ous*
6315	1025618	*-al*

```
barplot(c(557, 1410, 6315),
        ylab = "type frequency",
        names.arg = c("-ish", "-ous", "-al"))
```



The discrepency in type counts is already much lower and when we take the ratio, we can see that all of a sudden *-ish* is far ahead. It might not be as frequent as the other two suffixes, but it is also relatively more evenly spread over different roots. You could also imagine the other suffixes being bound to fewer roots that are very high in frequency individually.

```
barplot(c(557, 1410, 6315) / c(8196, 168294, 1025618),
        ylab = "type-token ratio",
        names.arg = c("-ish", "-ous", "-al"))
```

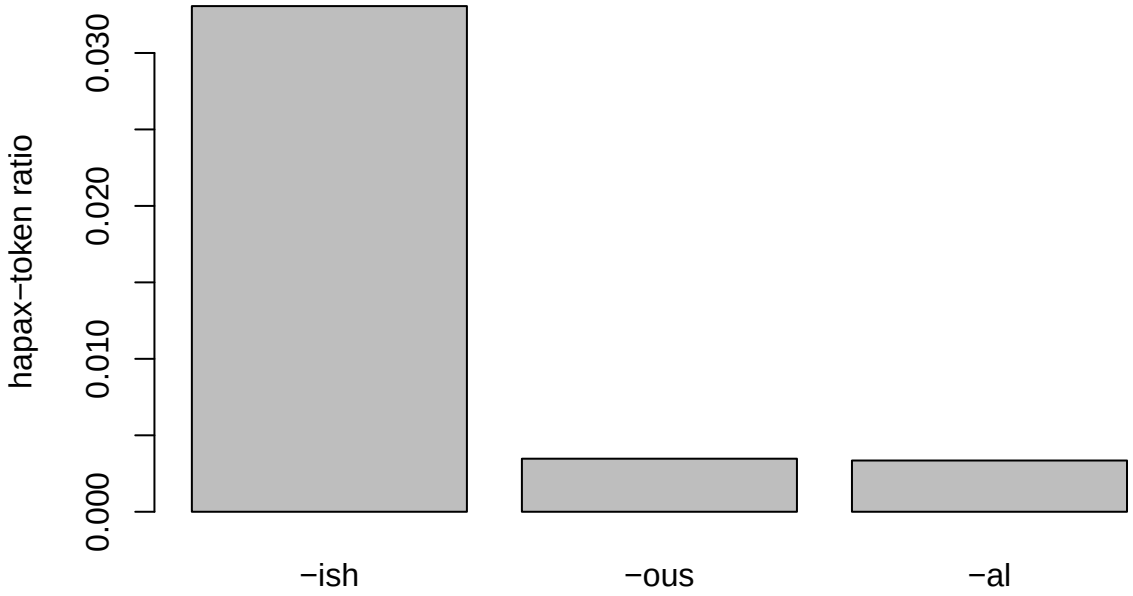


An even better measure is the hapax-token ratio. Instead of measuring the spread over different roots, we can measure the amounts of roots that only occur once. This may serve as an indicator for the formation of neologism. To get hapax frequency, we can add the following to our queries: `[... & f(word) = 1]`. We can observe the following result:

hapaxes:	types:	tokens:	suffix
271	557	8196	*-ish*

hapaxes:	types:	tokens:	suffix
584	168294		*-ous*
3438	1025618		*-al*

```
barplot(c(271, 584, 3438) / c(8196, 168294, 1025618),
        ylab = "hapax-token ratio",
        names.arg = c("-ish", "-ous", "-al"))
```



Even though *-ish* is comparatively infrequent, it occurs with a lot of types and seems to be used to form neologisms more often than the other two. This might indicate that it is more productive, and also that it is a more recent phenomenon because none of the uses have had the chance yet to become frequent individually. Another observation we can make is that while *-al*'s type-token ratio is higher than that of *-ous*, the hapax-token ratios are almost even. The data suggests that their productivity is similar, but that *-al* is more wide-spread in the lexicon, which makes intuitive sense.

10.1 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. ;)

Appendix

Term paper

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.

Any information about registration and deadlines to be announced ...

What makes a good term paper

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. You are free to pick any linguistic phenomenon as long as you demonstrate that you understood and

are able to apply the empirical techniques we have introduced in the lecture and in the seminar. In general, the more specific you are about it the better.

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.

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.

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[9]

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

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

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.

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.

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
```

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

```
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:
```

Academic posters

10.1.1 General information

Instead of student presentations on a given topic, you have the opportunity to investigate a phenomenon of your choice during one of two project days. Your topic serves as a first attempt at picking a linguistic phenomenon and developing a research question about it. It can but does not have to serve as a starting point for your term paper. Visit the section about the term papers for more information.

Your poster presentation 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.[10]

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.

10.1.2 Poster session

Project day format:

- Posters and Video contribution will be available in a repository and subsequently uploaded to Blackboard
- Presentations will be split up into 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 your contribution at least 24h before the presentations
- submit a **.pdf** file or in an **image format** (.bmp, .jpg, .svg,...), **not** .pptx or similar

Command line

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.

10.1.3 The shell

The shell is what is running in your terminal and interprets your commands. It is a bit like the counterpart of the background process of Windows, Linux or macOS that draws windows. You can move through your directories, create and delete files, run programs and much more. For example, instead of moving your mouse over your browser icon (let's say Firefox) and double clicking it, you would type the name of the program `firefox` and hit enter. `cqp` for example is a program that has a text-based interface, so if you type `cqp` in your terminal, you are opening

[10] 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.

in a different program. Some shell commands run programs that only have text output and no interface, such as `wc`. If you type `wc -l` and a filename `wc -l myfile.txt`, you run a program called `wc` which stands for 'word count' and you run it with the option `-w` which stands for 'word' and ask it to run on `myfile.txt`. It outputs the number of lines your file has. If you skip `-l`, you get the word count by default. In the end a shell is just a way to run programs like a graphical desktop is. The difference is that it is an interface which has full scripting ability.

Let's say I have got a frequency list `freqs.txt`, and I need to know how many types of words are in there. This is not an in-built function inside `cqp`, but (or rather because) you can easily do that in your shell like so: `wc -l freq.txt` Your freq list has one word per line so the line count is your type frequency. What if you left some words in there, that you should have excluded? You could either rerun your `cqp`-queries and get a new data set or simply use shell commands to clean up your data. If you want to remove every word ending in *ing* you could do this: `grep -v ".*ing" myfile.txt | wc -l`. Here, you use the program `grep` to search for a regular expression (those should look familiar) with the `-v` option for invert to get everything excluding your search term. Then we 'pipe' it into `wc -l` from before. That means we pass the result of `grep` into `wc`. This time we should get our type frequency without words ending in *ing*.

10.1.4 Copy & Paste

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

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

10.1.5 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.

10.1.6 Easterеggs

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

References

Anderwald, Lieselotte. 2011. Norm vs variation in British English irregular verbs: The case of past tense *sang* vs *sung*. *English Language & Linguistics*. Cambridge University Press 15(1).

85–112.

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

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.

Horsmann, Tobias, Nicolai Erbs & Torsten Zesch. 2015. Fast or accurate? – a comparative evaluation of PoS tagging models. *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.

Kaunisto, Mark. 1999. Electric/electrical and classic/classical: Variation between the suffixes *-ic* and *-ical*. *English Studies*. Taylor & Francis 80(4). 343–370.

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.

Plag, Ingo, Christiane Dalton-Puffer & Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics*. Cambridge University Press 3(2). 209–228.

Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive. In Günther Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 379–411. Mouton de Gruyter.

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

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

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

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.