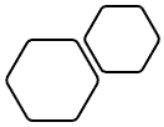
Rational communication across languages



An *Hauptseminar* brought to you by

Luigi Talamo

Department of Language Science and Technology

UNIVERSITÄT DES SAARLANDES

Credits

The present course is loosely based in its structure and aims on dr. Degaetano-Ortlieb's previous advanced seminars on rational communication. See for instance:

https://et-wiki.lst.uni-saarland.de/unserwiki/doku.php?id=teaching:ws_2019-20:hs_rational_communication

We give credits to Stefania and her excellent Hauptseminar on information-theory perspectives on language variation.

Motivation

- Languages are wonderful **tools to communicate**, most likely unique to human beings.
- There are about **6000-8000 languages in the world** (depending on how you count them).
- Why languages are **so diverse**? Is this **variability functional** for communication?
- Are there some languages more **expressive** than others? And easier to **learn**?

The faculty of language seems to be a prerogative of human beings. Some scholars, most notably, Chomsky, thinks that the ambiguity of languages makes them unsuitable as communication tools: in this perspective, languages serve to the primary function of thinking, communication being a sort of by-product. However, the ambiguity of languages can be solved by analyzing them in context i.e., taking into account the situation(s) in which linguistic utterances are produced. From this perspective, we will analyze languages for what we consider their primary function i.e., communicating something in a given context.

Our wonderful tool is multiplied for about 6000-8000 times, giving rise to the biblical Babel tower; only very few skilled and devoted persons can master dozen of languages and most of the time we have to rely on so-called *linguae francae* i.e., global languages such as English, French and Spanish. But why and how languages are so dramatically different, if they serve the common purpose of communication and the cognitive faculties are the same for all human beings? The reason is simple: culture and historical matters. Languages are human tools and like other human tools such as hammers, roads and spaceships are different implementations of the same blueprint.

Setting apart offensive ideas such as 'the language X is better/worser/more musical/less cacophonic than language Y', we can ask ourselves whether *some*

languages are more easier to learn than others, or more expressive than other with respect to certain domains of meanings or functions. Again, this is probably rooted in human cultures. When it comes to learnability, global languages, standardized languages and pidgins are easier to learn simply because they are evolved to be spoken by large human communities. As for expressivity, think of the classic and controversial example of Eskimo's several words for snow; despite being exaggerated by journalists and non-scholars, it is indeed true that Inuit (and not Eskimo, which doesn't exist as a single language) languages have more unrelated roots as well as derivatives for snow than English, indicating that snowy landscapes perhaps require more words to communicate snow-related things. For instance, *qanik* snow falling *aputi* snow on the ground *pukak* crystalline snow on the ground (see https://www.thecanadianencyclopedia.ca/en/article/inuktitut-words-for-snow-and-ice).

Motivation

A convenient way **to measure the amount of information** encoded in languages is offered by **information-theoretic metrics**, which use **computational models** of languages to represent **linguistic corpora** in order to quantify the complexity of **linguistic structures across languages.**

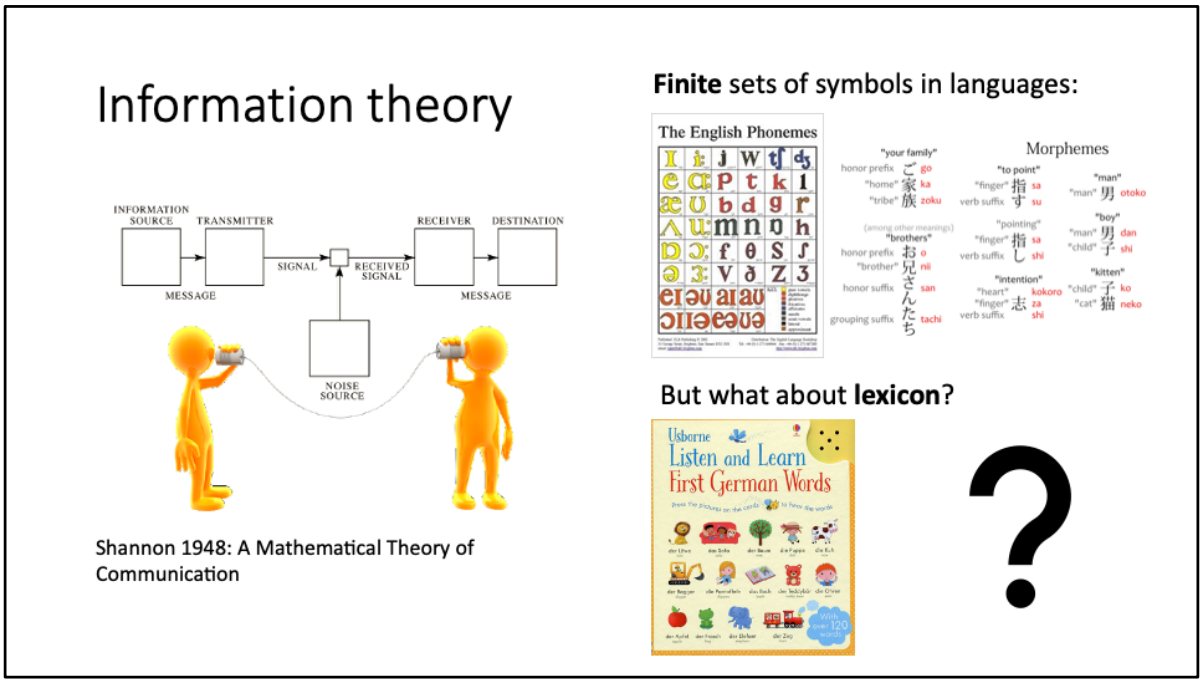
Our seminar thus lays at the crossroads of different disciplines, namely:

- Information theory
- Computational linguistics
- Corpus linguistics
- Linguistic Typology

Besides cultural and historical reasons for language complexity (learnability and expressivity), we may want to find an objective way to measure this complexity. Metrics of information i.e., how much information is carried by a given communication systems (tools) are offered by the field of information theory. Computational linguistics offers us model of representation for natural languages at different levels of analysis, while corpus linguistics deals with the study of languages from a usage-based and data-driven perspective, allowing us to handle actual linguistic data and build up bottom-up theories on language. Since the focus of this seminar is cross-linguistic, most of the proposed papers will focus on 'classic' problems from LT, such as how words are arranged at the sentence/phrase level, why some morphological systems are more complex than others, what does constitute a word, …

Information theory: assumptions

- Human languages are **rational communication systems** i.e., they are optimized to **carry information** (expressivity) while at the same they keep the **communicative effort reasonable** (learnability).
- This **rationality** is pervasive in languages, **acting at every level of linguistic structure** (from phonetics to the lexicon).



The founding paper of Information Theory (Shannon 1948), which you will see quoted in every paper of our reading list, deals with tools such as the teletype (a sort of ancient fax) and the telegraph. However, this can be easily extended to all tools:

- having a message transmitted from an information source i.e., the transmitter to a destination i.e., the receiver, using a signal transmitted over a channel. The channel can be more or less noisy i.e., disturbance over a telephone call, noisy environment, distances, …
- using a finite set of symbols: several aspects of languages can be approximated to a finite set of symbols. For instance, we have a finite set of phonemes at the phonetic levels, graphemes i.e., symbols/letters/ideograms at the ortographic levels, morphemes at the derivational or inflectional levels. However, words are not a finite set, as languages possess the so-called infinite productivity i.e., we can invent new words from existing ones. We'll see how information-theoretic measures deal with potentially-infinite set of symbols.

Information theory: Shannon's entropy

The amount of information of a given system (its **informativity**) is quantified in terms of **entropy**, expressing the degree of uncertainty:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where *n* is the size of the set of possible items and *P* the probability of each item composing the data set.

Given a (more or less) finite set of items and a message (system) transmitted over a channel, we can measure the quantity of information contained in the message by using entropic measures. Entropy is based on the observed probabilities of the items composing the message (the data set) and is expressed as the sum of the product of the probability of each item multiplied by the inverse logarithm of the probability of each item.

Information theory: Shannon's entropy

For instance, let's take the first class of nominal declension in Latin (N=12):

Case	Sg	Pl
Nom	Ros-a	Rosa-ae
Gen	Ros-ae	Ros-ārum
Dat	Rosa-ae	Ros-īs
Acc	Ros-am	Ros-ās
Voc	Ros-a	Ros-ae
Abl	Ros-ā	Ros-īs

- **How much is the entropy of the Latin inflectional system, restricted here to the first nominal class?**

Let's begin by counting the different realization for each cell i.e., how many **types** of inflectional endings we have for twelve total tokens (N) ?

Let's take the first class of Latin nominal declension, the very famous rosa-rosae-rosae. There are twelve possible cells in the paradigm (six cases X 2 numbers), but four endings are repeated across the paradigm, with the following distribution:

-a = 2
–ae = 4
–am = 1
–ā (long!) = 1
–ārum = 1
-īs = 2
-as = 1

We have then seven different types of inflectional endings: this is *n* of the sum, representing the size of our set.

Information theory: Shannon's entropy

We now assign probability values to the inflectional endings:

Case	Sg	Pl
Nom	Ros-a 0.17	Rosa-ae 0.33
Gen	Ros-ae 0.33	Ros-ārum 0.08
Dat	Ros-ae 0.33	Ros-īs 0.17
Acc	Ros-am 0.08	Ros-ās 0.08
Voc	Ros-a 0.17	Ros-ae 0.33
Abl	Ros-ā 0.08	Ros-īs 0.17

$$p(item) = \frac{freq(type)}{N}$$

and calculate the entropy as:

$$H = -(0.17 \times \log_2(0.17) + 0.33 \times \log_2(0.33) + 0.08 \times \log_2(0.08) + 0.08 \times \log_2(0.08) + 0.08 \times \log_2(0.08,2) + 0.17 \times \log_2(0.17)) = 2.56$$

- the amount of information for the first nominal declension of Latin is 2.56 bits

Probabilities are given by the simple equation: number of tokens for each type/ total number of tokens (maximum likelihood, or ML).
Recall the frequency of types and the total number of tokens = 12

-a = 2
–ae = 4
–am = 1
–ā (long!) = 1
–ārum = 1
-īs = 2
-as = 1

So, the amount of information for the first nominal declension of Latin is quantified in **2.56 bits**. We can compare this value to other nominal paradigms in Latin or in other languages.
But we can also quantify the entropy for a given inflectional ending and compare it with values from other endings.

Information Theory: Surprisal

A related entropic measure is **Surprisal**, which is focused on single items in a **given context**:

- Word in phrases, sentences, texts, collection of texts;
- Phonemes in phonetic inventories;
- inflectional endings in paradigms;
- …

$$S(x) = -\log_2 p(x)$$

How much are we **surprised** to find the item *x*, given **its probability in context**?

information content, **self-information**, **surprisal**, or **Shannon information (Wikipedia)**

Information theory: Surprisal

Case	Sg	Pl
Nom	Ros-a 0.17	Rosa-ae 0.33
Gen	Ros-ae 0.33	Ros-ārum 0.08
Dat	Ros-ae 0.33	Ros-īs 0.17
Acc	Ros-am 0.08	Ros-ās 0.08
Voc	Ros-a 0.17	Ros-ae 0.33
Abl	Ros-ā 0.08	Ros-īs 0.17

- What is the surprisal of the vocative plural?

$$S(voc.pl) = -\log_2 p(voc.pl|infl.paradigm) = 1.59$$

- And of the singular accusative?

$$S(acc.sg) = -\log_2 p(acc.sg|infl.paradigm) = 3.64$$

The inflectional endings of vocative plural, as well as other cases that use –ae is less difficult to learn/memorize in terms of information of the singular accusative, or other cases that use inflectional endings only attested once in the paradigm.

Information theory: Local informativity vs. (General) informativity

- Surprisal is a theoretic measure providing the local informativity of an item in a **particular** context = **contextual informativity**
- **General informativity** refers to the entropy of an item found in **every** context = **informativity**
- **They can differ a lot!** For instance, /d/ in *sudden* has an high contextual informativity. However, if we measure the entropy of /d/ in every occurrence of *sudden* in the Buckeye corpus, we discover that /d/ has indeed a low general informativity (Cohen-Priva 2015:248-249)

Local = contextual informativity
General informativity = informativity
Cohen-Priva 2015 will be discussed on December, the 17th. It is a paper on phonetics, whose main claim is that phonetic segments with low entropy = low information are deleted or shortened. Hence [ˈsʌdən] (alone) vs. [səˌn] (across the corpus)

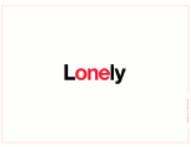
Information theory: Average Surprisal

Surprisal can be averaged on the **probabilities of the same item in different contexts**:

$$AvS(unit) = \frac{1}{|unit|} \sum_i -\log_2 p(unit_i|context_i)$$

where *unit* is the number of different contexts.

The best example here are probably **words**, which seldom occur alone…



A measure that take into account general informativity is average surprisal. We calculate the average of the surprisal for each different context.
Of course, it's really important what we take for unit, representing the window in which we observe the entropic measures. We will come back to the concept of unit when we'll approach computational models.

Information theory: Average Surprisal

Example: (Degaetano-Ortlieb & Teich 2019:13-14)

The word *book* compare in the following contexts, for a total number of 10 tokens (= times):

- **Jane reads the** *book*. Frequency: 8 times
- **Jane bought the** *book*. Frequency: 2 times

What is the surprisal of 'book' given the two contexts?

$$AvS(book) = \frac{1}{2}(-log_2p(0.8)) + (-log_2p(0.2)) = 1.32$$

Let's take the example discussed in Degaetano-Ortlieb and Teich 2019, which is the second paper of our reading list and is about the development of scientific English over nearly 4 centuries. Studies as such really need to take the context into account, in order to measure the surprisal of a given word in different contexts.
The surprisal of book given the first context is quite low, i.e. –log2(8/10) = 0.32 bit, while in the second context is high i.e., -log2(2/10) = 2.32 bits
What is the surprisal of book given the two contexts?

Information theory: some problems

In order to work as expected, **information-theoretic** measures should be applied to a data set meeting the following conditions:

- **finiteness**: a **finite set** of symbols (letters, phonemes, morphemes …words!); (cfr. slide no.6)
- **stationarity**: the probability of a preceding item **doesn't influence** the probability of a following item.

Tentative solutions:

- set a **minimum size for data**
- use **estimators**, which **approximates** the probability of words.

Bentz et al. (2017):4
Either one or both conditions are not met in language systems. The finiteness condition is easier met for phonological and morphological systems as well as for some syntactic constructions (the set of symbols is finite and you'll probably encounter all phonemes/morphemes/syntactic constructions in a decent corpus), but stationarity is hardly met as language symbols are repeated. This tends to overestimate the entropy of given items, as we assume that the data set we are studying is much more complex than it actually is.
In order to find a reasonable number of different words (=types) we will have to take corpora with more than 50K of tokens.
As for stationarity, you'll find that authors use estimators to 'correct' the probability of words: these methods are clearly beyond the scope of this course, it'll suffice to acknowledge their role. In order to overcome the probability problem, one paper uses the Kolmogorov complexity, which refers to individual objects and not to objects as member of a set (Geertzen et al. 2017:32-33).

Information theory: Relative Entropy (KLD)

- Useful when comparing **times slides in a diachronic corpus** or **translations of the same text**;
- How much does the entropy of the data set A fit the entropy of the data set B i.e., how much the **probability distributions** of the two data sets coincide?

Kullback-Leibler Divergence:

$$KLD(A||B) = \sum_i p(item_i|A) \log_2 \frac{p(item_i|A)}{p(item_i|B)}$$

If A = B, KLD is 0

When comparing different linguistic scenarios such as time slides in a diachronic corpus or translations of the same text, it's useful to have measure of relative entropy.
KLD compares the probability distributions of the data set A with the data set B by predicting how many additional bits are necessary 'to encode a given data set *A* when a (non-optimal) model based on a data set *B* is used (Degaetano-Ortlieb & Teich 2019:10).

Information theory: Relative Entropy (KLD)

Given the following original passage (along with probabilities), calculate the KLD for its German, Italian, French or English translations:

"Macondo (0.04) era (0.04) entonces (0.04) una (0.04) aldea (0.04) de (0.18) veinte (0.04) casas (0.04) de (0.18) barro (0.04) y (0.04) cañabrava (0.04) construidas (0.04) a (0.04) la (0.04) orilla (0.04) de (0.18) un (0.04) río (0.04) de (0.18) aguas (0.04) diáfanas (0.04)" (García Marquez, Cien años de Soledad)

Poll: Which language do we choose?

Information theory: Conditional entropy

- How much does the entropy of an item *c* **influence** the entropy of another item *x*?

$$H(x|c) = -\sum_{x}^{n}\sum_{c}^{n} P(x,c)\log_2 P(x|c)$$

- Useful when *x* and *c* are **members of the same set** e.g., forms of inflectional paradigms, collocations, …
- However, the influence is **purely statistical** i.e., it does not depend on the form/meaning of the items

We have already seen that linguistic items are hardly met alone, thus constituting a problem for information theoretic measures intended for stationarity systems. We can take a further step, and ask how much the entropy of the item c conditions the entropy of the item x. It is particularly useful in paradigmatic systems, in which the user (speaker) has to choose between a datasets in which items are members. Since it's just statistics, it doesn't take into account the fact that some forms are easier to connect than others, simply because they show some resemblance i.e., the dat.sg. = gen.sg or singular forms have all ending vowels save the accusative, …

Information theory: Conditional entropy

Case	Sg	Pl
Nom	Ros-a 0.17	Rosa-ae 0.33
Gen	Ros-ae 0.33	Ros-ārum 0.08
Dat	Ros-ae 0.33	Ros-īs 0.17
Acc	Ros-am 0.08	Ros-ās 0.08
Voc	Ros-a 0.17	Ros-ae 0.33
Abl	Ros-ā 0.08	Ros-īs 0.17

- How much information we need to know the accusative singular, **if we already know the genitive plural**?

$$H(acc.sg.|gen.pl) = -(0.08 \times \log_2(0.08) + 0.08 \times \log_2(0.08) = 0.58$$

- And if we know the **dative singular** (= more likely?)

$$H(dat.sg.|acc.sg.) = -(0.08 \times \log_2(0.08) + 0.33 \times \log_2(0.33) = 0.81$$

If we already know a less likely ending, say, the genitive plural, it will take us little effort to fill the paradigm: just need 0.58 bits for the accusative singular. Taken alone, the dative singular requires 1.59 bits (remember the Kinder-Überraschung slide).
On the other hand, knowing more likely endings, such as the dative singular, require more effort to know the dative singular, i.e., 0.81 bits

Computational linguistics: models

Language model = space in which we can observe the probabilities of our set

Different **levels of analysis** requires different **languages models**:

- Lexicon: we need to capture the links between **words**, so we will use string of words, or **n-gram based models**:

"The quick brown fox jumps over the lazy dog"

Unigram: the, quick, brown, fox, jumps, …

Bigram: the quick, quick brown, brown fox, …

Trigram: the quick brown, quick brown fox,

Four-gram: the quick brown fox, quick brown fox jumps

….: the limit is somewhat set to **five-gram**

In order to use our information-theoretic measures we need to represent the portion of language we are studying i.e. the corpus with a model. This is necessary as we have to set a space in which observe the probabilities of our set. For instance, the Macondo example from the last slide uses the unigram model.

Computational linguistics: models

N-gram based models are also good for other levels in which items are found in short-range contexts, such as:

Phonetics: **strings of phones**

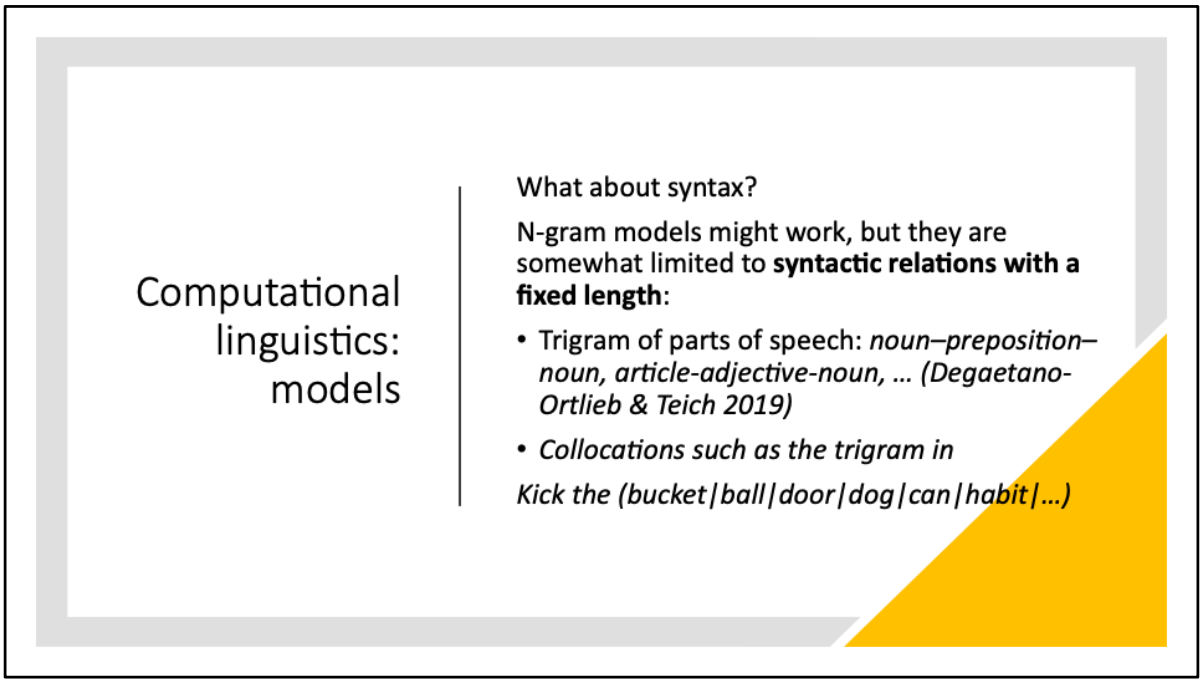
[no͡ʊt͡ɪs] [noʊəs]

uniphone: n, o, ʊ, t or zero,…

biphone: no, oʊ, ʊt or oʊ, …

triphone: noʊ, o͡ʊt, ʊ͡tɪ or ʊəs, …

(Cohen-Priva 2015)



N-gram viewer at Google: https://books.google.com/ngrams

Computational linguistics: models

Summing up

N-gram: gram in a context of N-1 gram

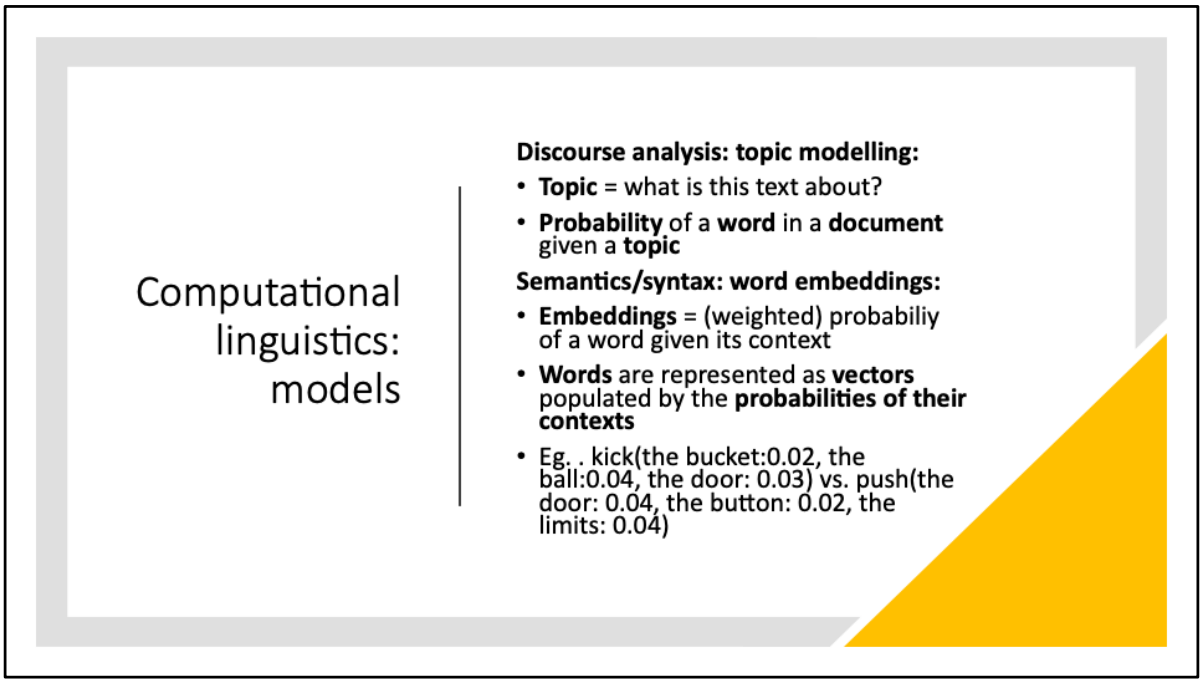
unigram: just the gram itself

Bigram: the gram plus one preceding gram

Trigram: the gram plus two preceding gram

...

Good for lexicon, phonetics, morphology. **Problematic** for syntax.



If we want to move the discourse analysis, we'd better employ a computational model taking into account a topic as wells. Topics are usually very general, for instance in scientific languages can be names of disciplines: Physiology, Chemistry, Geography, …

Vector-based models are good for doing semantics and, to some extent, investigate problems from syntax. A vector-based model is built by taking the probabilities of words in different context. For instance, let's take the verb 'kick' and 'push', which are represented by vectors populated by the probabilities of 'the bucket', 'the ball', 'the door' and so on. We can for instance infer their similarity (to a certain extent…) as they share at least one context, 'the door'.

These two models are used in Bizzoni et al. 2020.