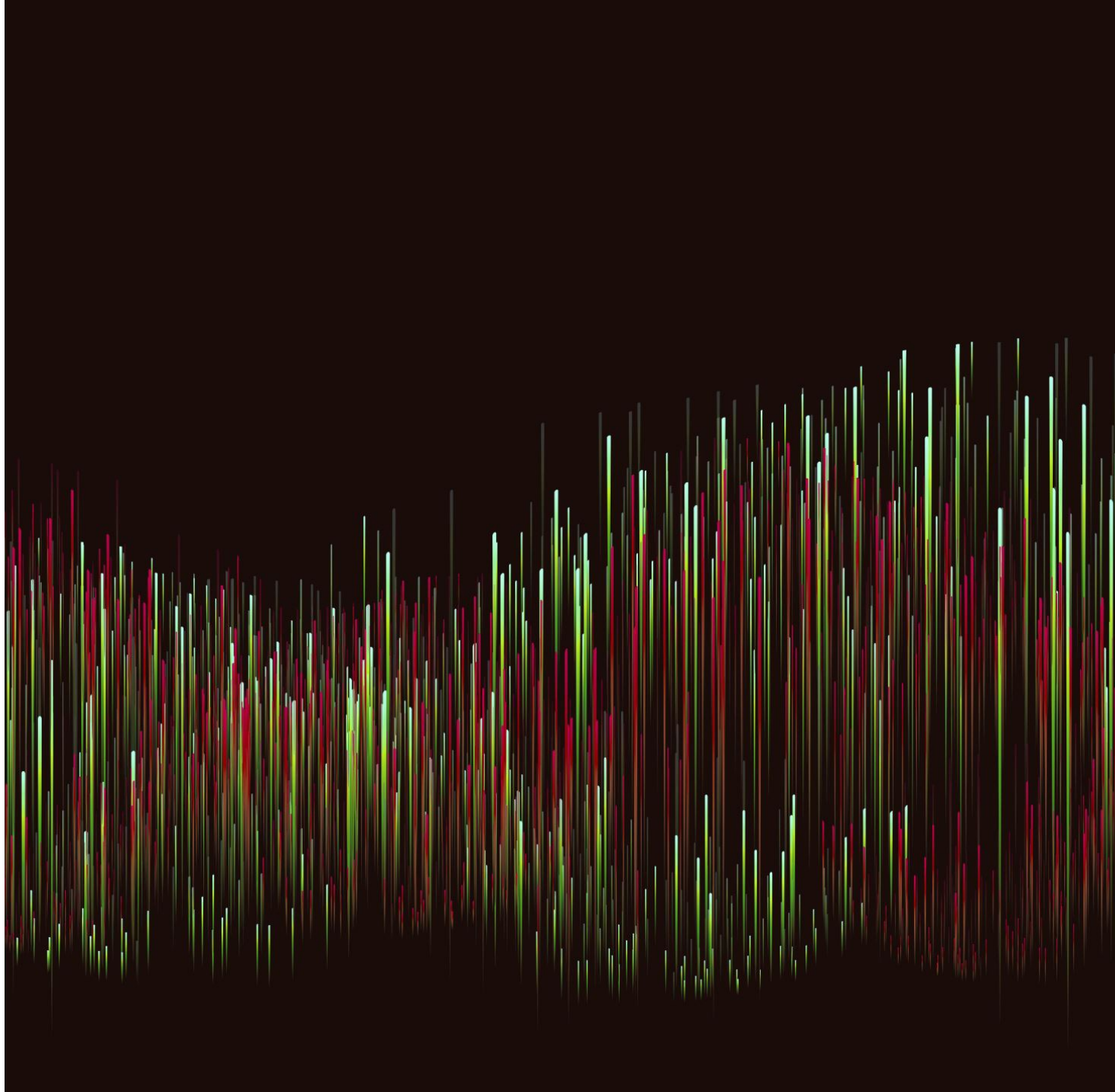
Linguística de *Corpus*

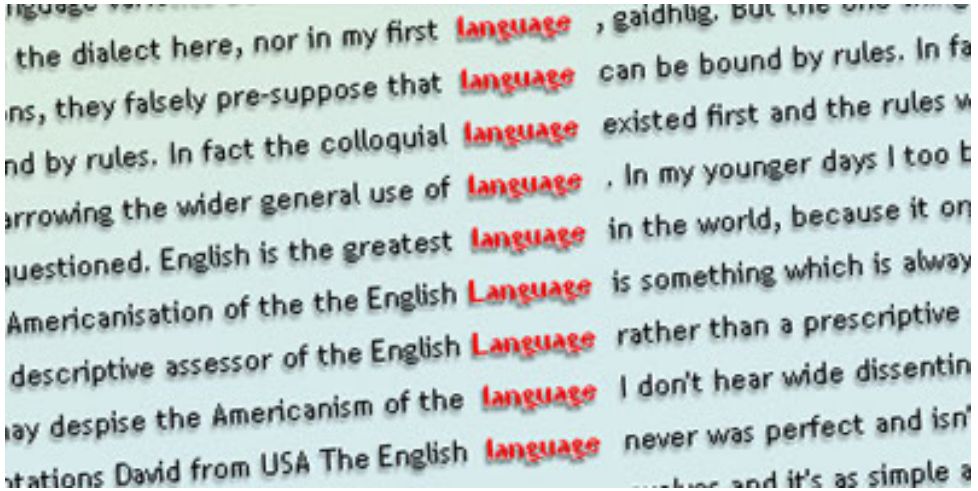


CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Linguística de *Corpus* e dados

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- O que é um *corpus*?
- O que trouxeram os *corpora* para a Linguística?
- O que é a Linguística de *corpus*?
- Dados e métodos da Linguística de *Corpus*
- Usos da Linguística de *Corpus*



O que é um *corpus*?

"A **corpus** is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language."

"Words such as collection and archive refer to sets of texts that do not need to be ordered, or the selection and/or ordering do not need to be on linguistic criteria. They are therefore quite unlike corpora."

(Sinclair 1996: 4,5)

O que é um *corpus*?

"A **corpus** is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language." (Sinclair 1996: 4,5)

dados linguísticos autênticos

crtérios de seleção explícitos

amostras da língua

O que trouxeram os *corpora* à Linguística?

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Dados autênticos, refletindo tudo aquilo que é produzido (por oposição a conceções individuais da língua/exemplos fabricados pelos linguistas)

- Dados quantitativos sobre fenómenos linguísticos – o que é raro, o que é frequente, o que coocorre sempre, o que nunca coocorre…

- Sendo uma amostra, um *corpus* permite a quantificação e a extrapolação dos resultados (como qualquer outro estudo que use amostragem)

O que trouxeram os *corpora* à Linguística?

- Uma nova perspetiva teórica sobre o sistema linguístico

"Analysis of extended naturally occurring texts, spoken and written, and, in particular, computer processing of texts have revealed quite unsuspected patterns of language... [The] contrast exposed between the impressions of language detail noted by people, and the evidence compiled objectively from texts is huge and systematic ...

The language looks different when you **look at a lot of it at once** . . . . "

(Sinclair 2004: 1)

O que trouxeram os *corpora* à Linguística?

- Uma nova perspetiva teórica sobre o sistema linguístico

Princípio da idiomaticidade (Sinclair 1991, 2002): a maioria dos textos é composta essencialmente por expressões de mais que uma palavra que não ocorrem por acaso. Estas são opções/escolhas únicas, i.e., o que vem antes e depois de uma dada palavra/expressão é condicionado pelo idioma – os dados da língua real – a que os falantes estão expostos.

O que trouxeram os *corpora* à Linguística?

- Uma nova perspetiva teórica sobre o sistema linguístico

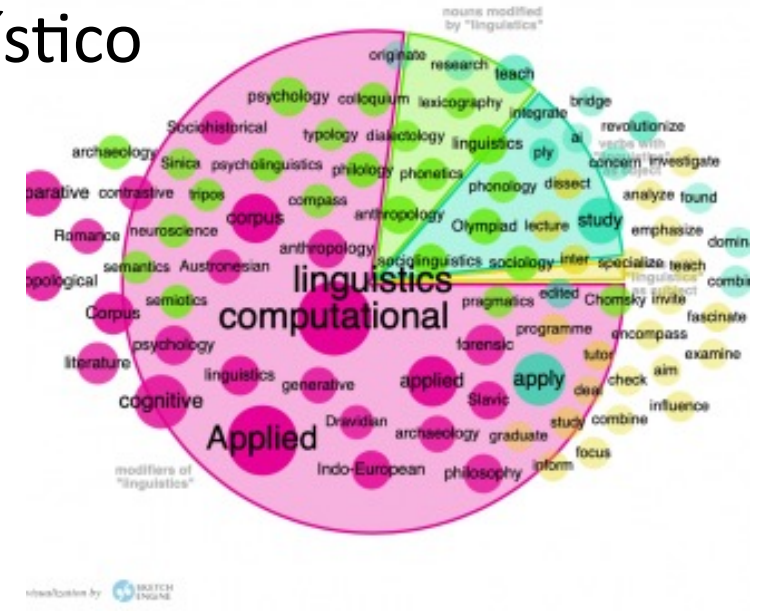
hard/soft drugs	vs.	?# *hard/soft cigar...*
strong tea	vs.	?#*powerful tea*
sour milk	but	*rotten egg*
phone booth	but	*water closet*
machine translation	not	?*computer translation*

(Sinclair 1991, 2002)

O que trouxeram os *corpora* à Linguística?

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Novos dados
- Uma nova perspetiva teórica sobre o sistema linguístico
- Novos métodos e ferramentas



O que é a Linguística de *Corpus*?

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- A Linguística de *Corpus* é uma subárea da Linguística que lida com grandes quantidades de dados linguísticos autênticos para estudar as línguas.

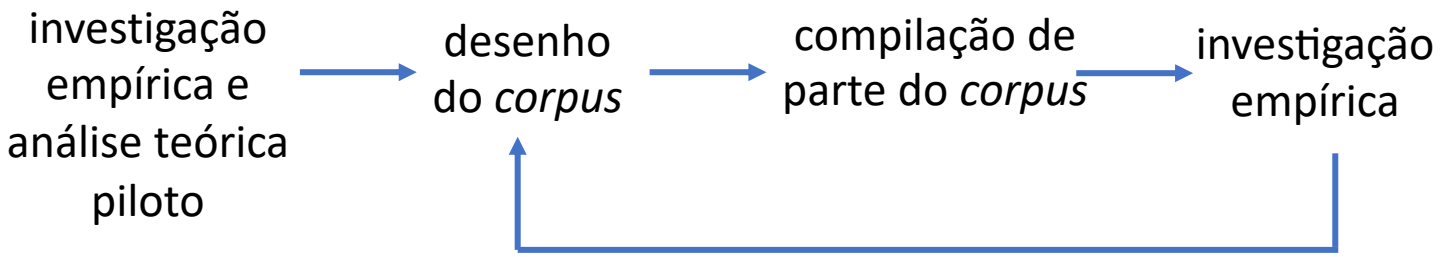
- Tem como base uma noção específica de *corpus.*
- Usa métodos computacionais.

→ Compreende métodos e questões específicos e desenvolve/usa ferramentas específicas.

Dados e métodos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Compilação e design do *corpus*



- Critérios de seleção
- Métodos de seleção
- Amostragem
- Normalização
- Codificação (*vs*. anotação)

Table 1 Situational parameters listed as hierarchical sampling strata

1. *Primary channel*. Written/spoken/scripted speech
2. *Format*. Published/not published (+ various formats within 'published')
3. *Setting*. Institutional/other public/private-personal
4. *Addressee*.
   (*a*) Plurality. Unenumerated/plural/individual/self
   (*b*) Presence (place and time). Present/absent
   (*c*) Interactiveness. None/little/extensive
   (*d*) Shared knowledge. General/specialized/personal
5. *Addressor*.
   (*a*) *Demographic variation*. Sex, age, occupation, etc.
   (*b*) *Acknowledgement*. Acknowledged individual/institution
6. *Factuality*. Factual-informational/intermediate or indeterminate/imaginative
7. *Purposes*. Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship, . . .
8. *Topics*. . . .

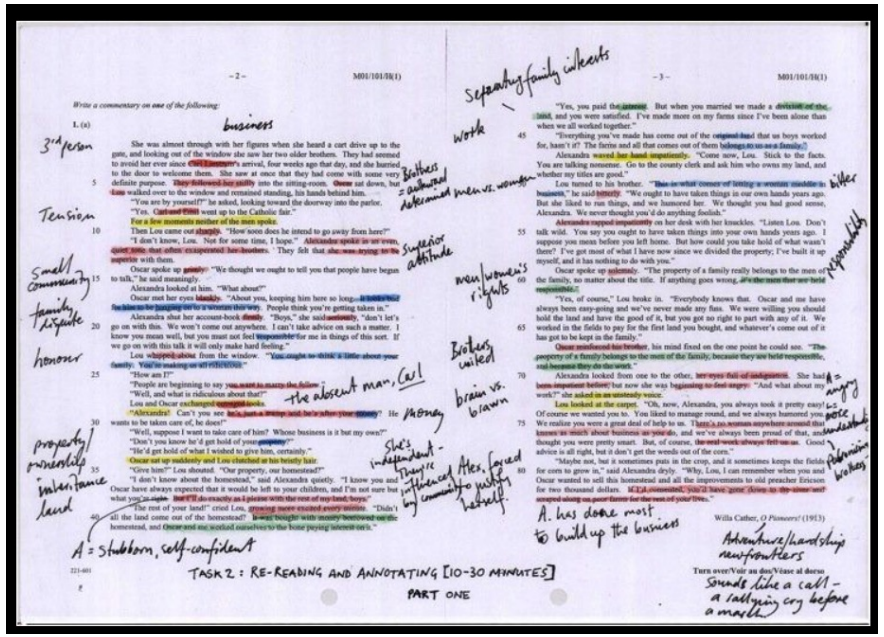
Dados e métodos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Anotação de *corpus*

*Corpus annotation — **adding interpretive information into a collection of texts**—is valuable for a number of reasons, including the validation of theories of textual phenomena and the creation of corpora upon which automated learning algorithms can be trained."*

(Hovy & Lavid 2010)



Dados e métodos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Anotação de *corpus*/dados

i) Manual

+ exige menos preparação do *corpus*

+ é menos dispendiosa no que respeita a fenómenos complexos

- consome muito tempo

- origina pouca quantidade de resultados

ii) Automática

- grande investimento em sistemas de preparação e etiquetagem do corpus

- resultados de baixa qualidade

+ geração de grandes quantidades de dados

+ muito rápida

Dados e métodos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Anotação de *corpus*/dados -> ciência de anotação de dados

1. *Corpus* de treino (representatividade e diversidade)
2. *Tag set* de anotação e manual de anotação (indicações gerais e preconceitos)
3. Viabilidade da anotação do *corpus* de teste (tempo/esforço e adequação)
4. Avaliação de resultados (decisões dos anotadores) e medidas corretivas
5. Nível de concordância entre anotadores
6. Treino de ferramentas de aprendizagem automática (*corpus* de treino vs. teste)

Dados e métodos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Ferramentas & recursos para o tratamento de dados

- TEI – Text Encoding Initiative
- Concordanceiros (WordSmith, AntConc, SkecthEngine, …)
- POS taggers/ parsers…
- CQL - Corpus Query Language/CQP - Corpus Query Processor
- Anotadores (UAM Tools, MMAX2, …)

Dados e métodos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Métodos de análise

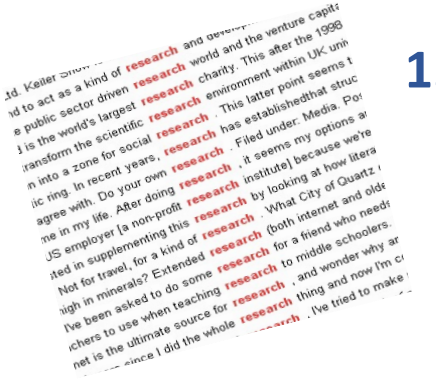
- Análise quantitativa
    - categorização/classificação dos dados
    - contagens
    - frequências
    - estatísticas

- Análise qualitativa
    - identificação de fenómenos (→ categorização)
    - análise detalhada e equitativa dos dados (raros e frequentes)

Usos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

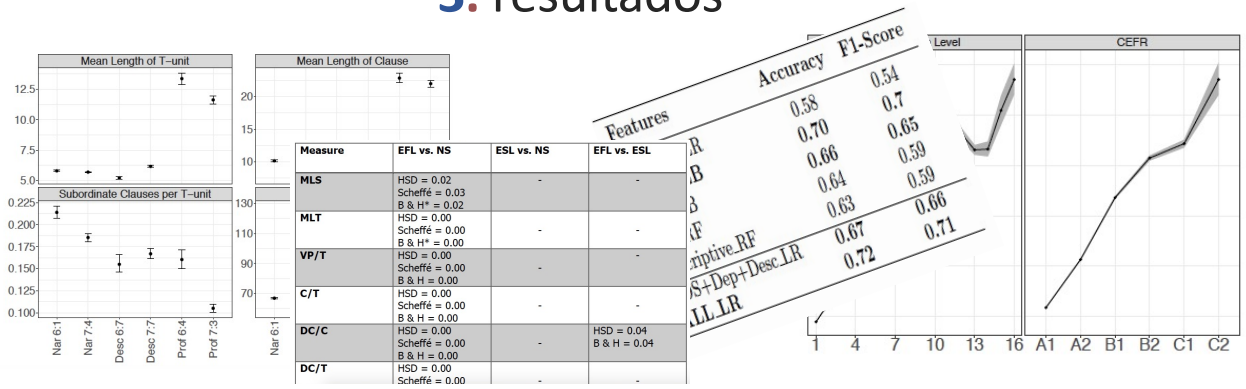
- **Aprendizagem Automática**



1. dados & anotação de dados

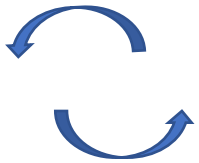
2. algoritmos (ex.: DeepLearning)

3. resultados



Measure	EFL vs. NS	ESL vs. NS	EFL vs. ESL
MLS	HSD = 0.02 Scheffé = 0.03 B & H* = 0.02	-	-
MLT	HSD = 0.00 Scheffé = 0.00 B & H* = 0.00	-	-
VP/T	HSD = 0.00 Scheffé = 0.00 B & H = 0.00	-	-
C/T	HSD = 0.00 Scheffé = 0.00 B & H = 0.00	-	-
DC/C	HSD = 0.00 Scheffé = 0.00 B & H = 0.00	-	HSD = 0.04 B & H = 0.04
DC/T	HSD = 0.00 Scheffé = 0.00	-	-

4. iteração



Usos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Estudos linguísticos: novos modos de comunicação

  G&T commenta (https://clunl.fcsh.unl.pt/en/investigacao/projetos-curso/gt-comenta/)

  NetLang (https://sites.google.com/site/projectnetlang/corpus)

- Ensino/Aprendizagem de línguas

  Corpus de PLE/PL2(http://teitok.clul.ul.pt/cople2/

  MultiMind (https://www.multilingualmind.eu/

Usos da Linguística de *Corpus*

- Estudos Literários

  Distant Reading for European Literary History (COST Action CA16204) (https://www.distant-reading.net/)

- Estudos Sociais: inclusão e integração

  (De)Othering (https://deothering.ces.uc.pt/en_GB/about/#overview_anchor

  Exprimi (https://clunl.fcsh.unl.pt/en/investigacao/projetos-curso/exprimi/)

  Inclusive Courts (https://inclusivecourts.pt/en/)

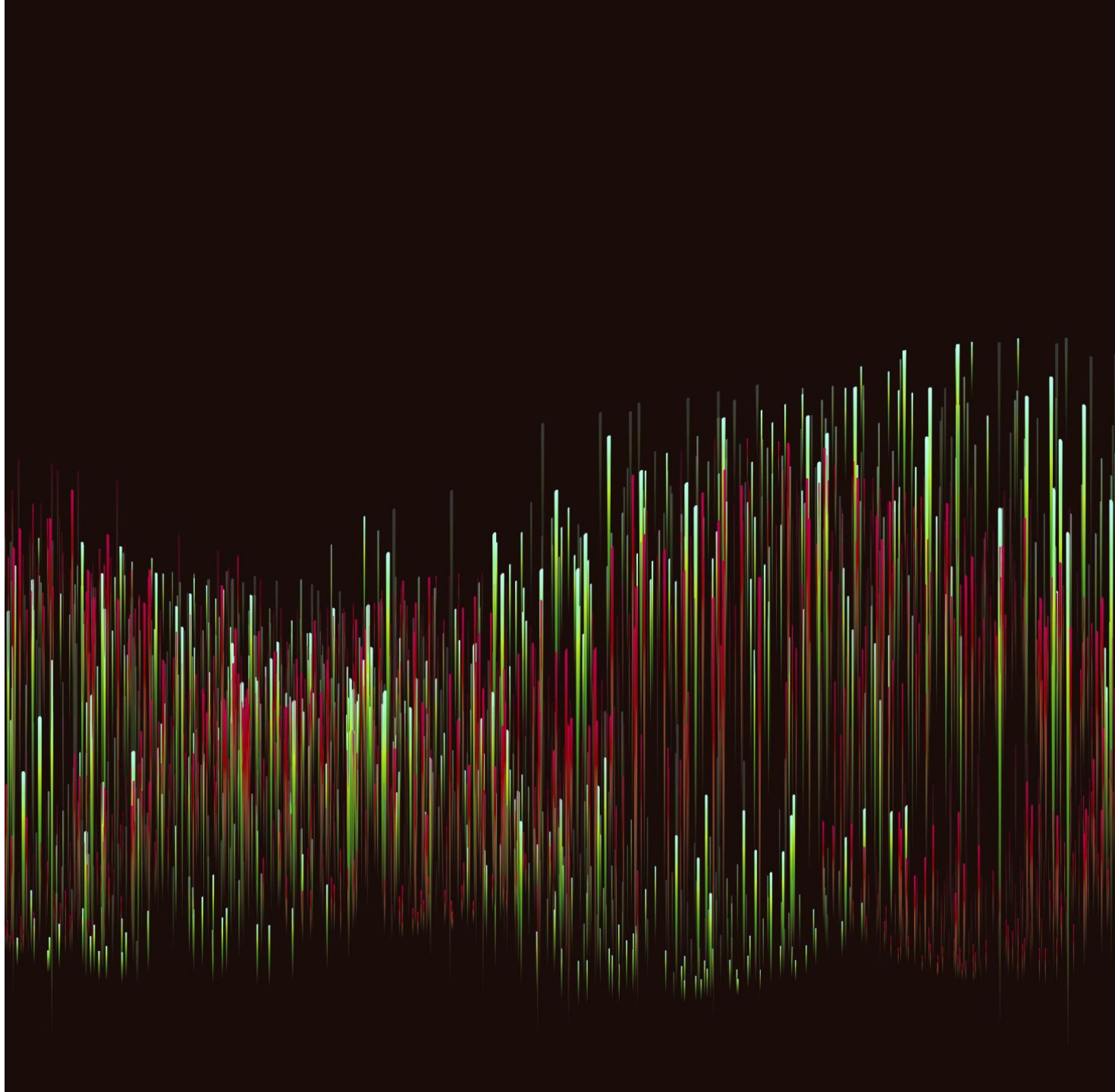
Usos da Linguística de *Corpus*

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

- Repositórios com *corpora*

https://portulanclarin.net/repository/search/

http://www.linguateca.pt/corpora_info.html



Raquel Amaro, Chiara Barbero, Sílvia Barbosa

CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Erasmus+

NOVAFCSH

EN	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
PT	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.