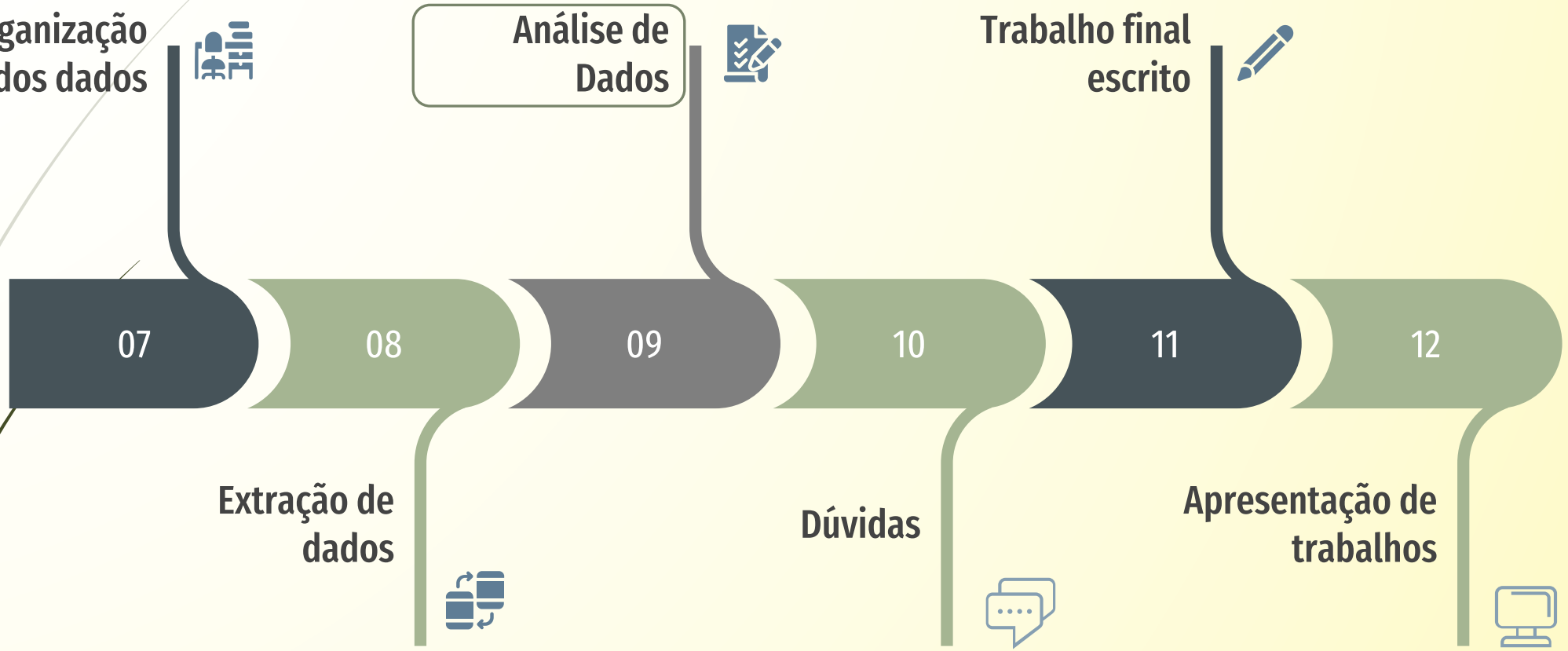
Humanities Going Digital (HUGOD)

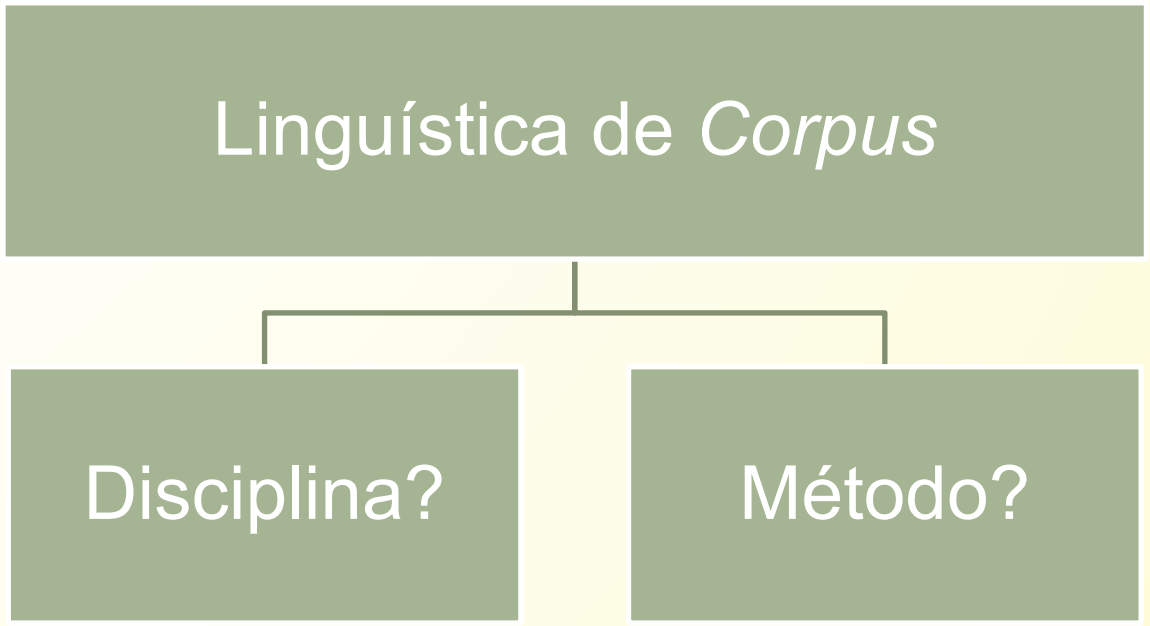
[2020-1-CZ01-KA226-HE-094363]

Análise de dados	Sessão 09	31/10/2022	Chiara Barbero Sílvia Barbosa

AULAS


Organização dos dados
Análise de Dados
Trabalho final escrito
07
08
09
10
11
12
Extração de dados
Dúvidas
Apresentação de trabalhos

Dos dados linguísticos à extração de informação específica



Linguística de *Corpus* – Disciplina

(Tognini-Bonelli, 2001)

- LC como perspetiva teórica que considera que as regras linguísticas baseiam-se no uso que os falantes fazem da língua (usage-based), e as mudanças na língua ocorrem na medida em que esta é usada para comunicar.
- Porque reúne atividades de teorização, recolha de dados, análise e argumentação.
- Porque conjuga análises quantitativas (estatísticas) e observações qualitativas sobre os fenómenos linguísticos.
- Por definir uma linha de investigação e uma abordagem filosófica nos estudos linguísticos.

Linguística de *Corpus* – *Método científico* (McEnery et al., 2006; Brezina, 2018)

- Por não se limitar a um único aspeto linguístico, mas pelo contrário, por ser aplicável a qualquer área de investigação linguística.
- Científico, por quê?
  - Porque utiliza dados empíricos – i.e. autênticos.
  - Garante a replicabilidade dos dados

    implica transparência nos critérios de seleção dos textos e na escolha das técnicas de análise de dados.
  - Minimiza subjetividades e prejuízos humanos para obter resultados o mais possível fidedignos.

Bacelar do Nascimento (2002)

- *"A nosso ver, os corpora não constituem, em si próprios, um novo ramo da linguística (…). Consideramos, antes, de uma forma abrangente, que os corpora proporcionam novas maneiras de estudar as línguas, das quais resultam descrições, generalizações e hipóteses teóricas de grande consistência porque fortemente enraizadas nos dados empíricos. (…) Por estas razões, achamos que um corpus se define não só por factores tão importantes como a sua dimensão, constituição, diversificação, estrutura e dinâmica de actualização, mas também, decididamente, pela variedade de utilizações que proporciona." (Bacelar do Nascimento 2002: 1-2)*

Diferentes Abordagens

Abordagens a partir de intuições ou de teorias linguísticas pré-existentes:

- Intuition-based
  - Procura as respostas (i.e. exemplos concretos) à intuições
- Corpus-based
  - Provar ou negar teorias previamente formuladas ou traços/características e construtos previamente identificados

Diferentes Abordagens

Abordagens top-down ou bottom-up:

- Corpus-based — análise dos dados por dedução/inferência
- Corpus-driven — análise indutiva a partir dos dados

Prós e contras das diferentes abordagens

- **corpus-based:** a escola teórica seguida pelo investigator pode influenciar a interpretação dos resultados (usados para para argumentar, testar ou exemplificar teorias já existentes)
- **intuition-based**: a intuição pode ser influenciada por múltiplos fatores que podem ser desviantes: e.g. variantes linguísticas; uso correto mas não prototípico; exemplos criados de raiz para motivar as intuições
- **corpus-driven**: cuidados com a abrangência dos dados - os *corpora* são repositórios de dados, mas não necessariamente exaustivos

Tipos de análises

Análises:

- quantitativa
  - Língua analisada em termos numéricos: frequências, colocações e coocorrências, estruturas /padrões, etc.
- qualitativa
  - Interpretação dos "números": porquê determinados fenómenos acontecem?

ESTAS ANÁLISES NÃO SE EXCLUEM MUTAMENTE, MAS SÃO COMPLEMENTARES!

Tipos de análises: exemplos

- **Análise contrastivas/comparativa**

- Estruturas sintáticas entre registos diferentes: (uso dos pronomes pessoais, uso das estruturas predicativas, etc...)
- Estudos multilingues em corpora paralelos (traduções de expressões idiomáticas)

- **Análise diacrónica**

- Evolução ortográfica de uma unidade lexical ou de um morfema

- Evolução/uso de uma expressão/forma

Tipos de análises: exemplos

- **Análise semântica**

- Valor semântico de uma categoria de unidades lexicais (ex. marcadores discursivos)
- Padrões de polissemias regulares

- **Análise lexical**

- Estudo das unidades multilexicais: caracterização, cristalização/variação
- Léxico de um domínio de especialidade

Perspetivas de análise: micro e macro

CLOSE READING

- Análise de um texto (romance, poesia etc.) individualmente – leitura atenta para caracterização da obra de acordo com, por exemplo, estratégias retóricas usadas pelo autor, simbolismos, referências culturais etc..

DISTANT READING

- Análise de um conjunto muito maior de dados - identificar tendências transversais a mais do que um texto para, por exemplo, questionar características de uma cânone literário, encontrar estruturas/léxico partilhados etc..

Objetivos

Usar pistas linguísticas para resolver questões de investigação:

a compilação de corpora e as ferramentas de tratamento permitem-nos aproveitar traços/padrões/particularidades linguísticas para encontrar evidências que respondam a questões de investigação maiores (tanto linguísticas, como de outra ordem)

- Exemplo: Como observar o nível de inclusividade da linguagem? Como compreender esse fenómeno?

Métodologia: podemos olhar para os nomes relativos às profissões/papéis institucionais? São usados maioritariamente na forma masculina ou feminina?

Referências bibliográficas

- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.
- Bacelar do Nascimento, M. Fernanda (2002) *O lugar do corpus na investigação linguística.* Centro de Linguística da Universidade de Lisboa
- McEnery, T., Xiao, R., Tono, Y., (2006) *Corpus-based language studies: An advanced resource book*. Taylor & Francis.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins

Avaliação

Bibliografia a consultar:

- https://repositorio.ul.pt/bitstream/10451/30696/1/Mendes_Lingu%C3%ADstica-de-Corpus-e-outros-usos-dos-corpora-em-lingu%C3%ADstica_draft_2016.pdf
- https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm

Para a próxima aula:

- Submeter o resumo
- Enviar as tarefas

Não esquecer: preencher os formulários no MOODLE (questionários e as tarefas)

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.