Scientific Writing in English: Techniques and Tools

Create your Own Corpus
Part V

Ethel Schuster, Ph.D.
Northern Essex Community College, USA
eschuster@necc.mass.edu

Your Corpus

- What is it?
- How to Annotate?
- Why create one?
- How to create one?
  - 9 Steps
- Format?
- How much, big?

Ethel Schuster, Ph.D.

Corpus

- Collection of text, sentences
  - based on papers
  - from your area
- With
  - Annotations

Annotations?

- Marks/notes made while reading text
- Can be
  - Underline
  - Highlight
  - Hand-written
  - “part-of-speech” tagging

Annotated Text

- Can help author
  - Construct an argument
  - Write a paper
  - Remember important facts

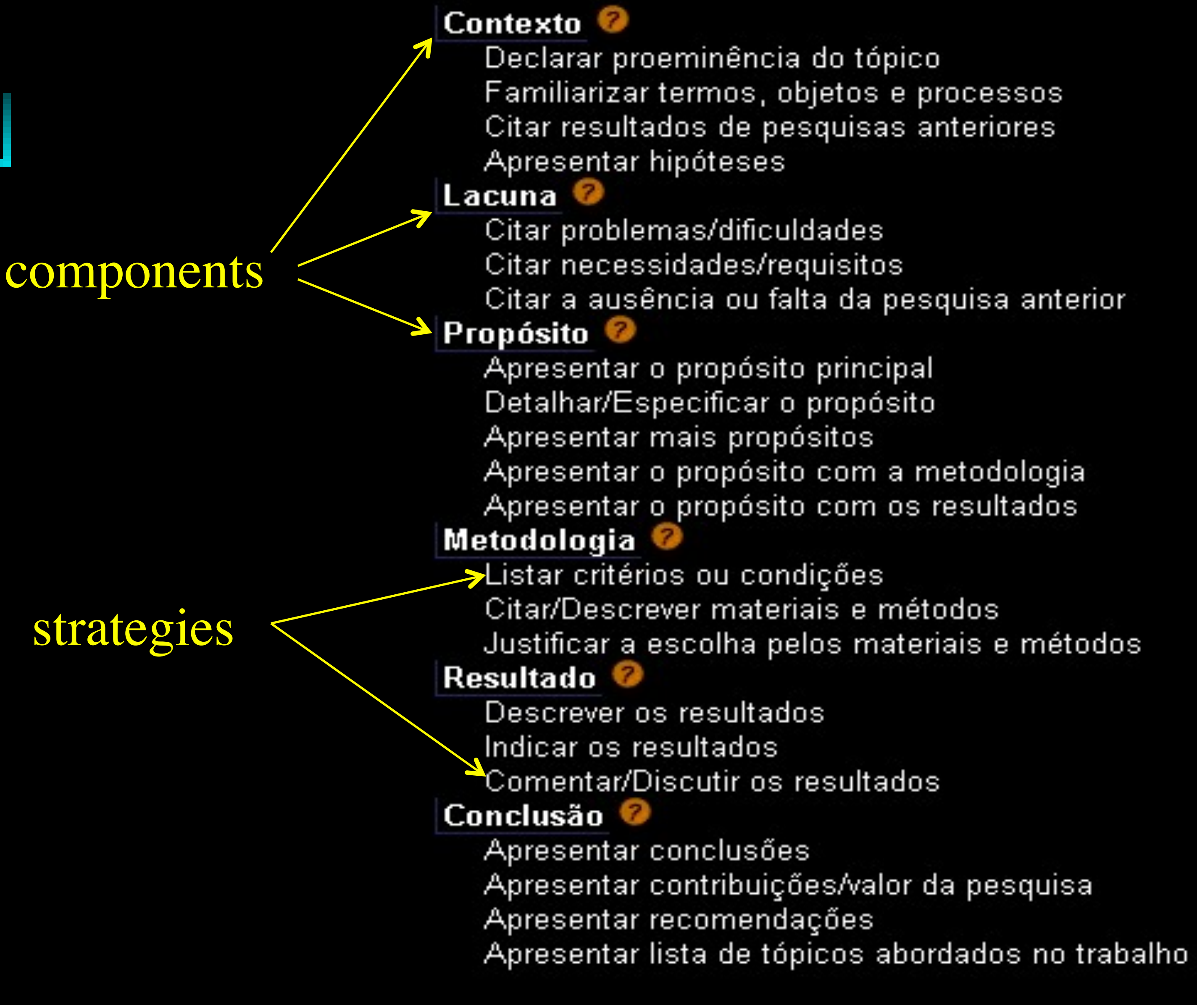
Components?

- Usually include
  1. Background/Context
  2. Gap
  3. Purpose
  4. Methods and Materials
  5. Results
  6. Conclusion

Strategies

- Components further classified into strategies, e.g.
  - Arguing about topic's prominence
  - Familiarizing terms or objects or processes
  - Listing criteria or conditions
  - Indicating/Describing materials or methods

Components and Strategies



components

strategies

Contexto
- Declarar proeminência do tópico
- Familiarizar termos, objetos e processos
- Citar resultados de pesquisas anteriores
- Apresentar hipóteses

Lacuna
- Citar problemas/dificuldades
- Citar necessidades/requisitos
- Citar a ausência ou falta da pesquisa anterior

Propósito
- Apresentar o propósito principal
- Detalhar/Especificar o propósito
- Apresentar mais propósitos
- Apresentar o propósito com a metodologia
- Apresentar o propósito com os resultados

Metodologia
- Listar critérios ou condições
- Citar/Descrever materiais e métodos
- Justificar a escolha pelos materiais e métodos

Resultado
- Descrever os resultados
- Indicar os resultados
- Comentar/Discutir os resultados

Conclusão
- Apresentar conclusões
- Apresentar contribuições/valor da pesquisa
- Apresentar recomendações
- Apresentar lista de tópicos abordados no trabalho

Scientific Writing Corpus

- Includes
  - Text
    - Annotated
    - With
      - Components + strategies

Before Starting to Annotate

- Read entire text
  - Important to have overview
- Do not try to understand research
- Focus on structure

Process of Annotation

- Determine meaning of sentence
- Interpret role given by author
- Focus on author's intentions
- Consider location and context of sentence
    - Must assign at least one label/ sentence

Categories and Labels

Categories	Labels
Background	<Background> … </Background>
Gap	<Gap> … </Gap>
Purpose	<Purpose> … </Purpose>
Methods	<Methods> … </Methods>
Results	<Results> … </Results>
Conclusions	<Conclusions> … </Conclusions>

Assigning Labels: Background

- Assessments of importance of topic, which justify studying it
- Claims that particular topic, technique, strategy, or method is important and deserves attention

Assigning Labels: Gap

- Some aspects of topic have not yet been studied or explored
- Little is known about subject
- Earlier attempts at studying it were unsuccessful or have produced conflicting
- Claims that something should be done

Assigning Labels: Purpose

- Goals/aims of research
- Hypothesis to be proven
- Objectives

Assigning Labels: Methods and Materials

- Description of data/method
- Methods discussed
- New method(s) proposed

Assigning Labels: Results

- Describe specifics
- Cannot be opinions, suggestions, or value judgments

Assigning Labels: Conclusions

- Overall conclusions of research
- Recommendations
- Suggestions, opinions

MAZEA for Annotation

- Tool for identification of components of abstract
- Only uses 6 of them: Background, Gap, Purpose, Methods, Results, Conclusion
- Based on machine-learning

MAZEA

- http://www.nilc.icmc.usp.br/mazea-web/

How to create corpus?

- Nine Steps
- Each step
  - Description
  - Explanation
  - Illustrated
    - One paper as example

Step 1

(a) Select well-written texts from reliable sources and produced by native speakers

(b) Read the material critically, annotating expressions that convey important messages and may be useful [to reuse] in the future

Step 1a Deconstructed

- Select paper from IEEE Computer: “Social Networking”
- Among most popular

Example: IEEE Xplore

“Social Networking”

By Weaver, A.C. And Morrison, B.B. [Univ. of Virginia, Charlottesville]

Computer Volume: 41, Issue: 2
Publication Year: 2013 , Page(s): 97-100

http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=4454412&contentType=Journals+%26+Magazines&punumber%3D

Citation and Abstract Downloaded

- Weaver, A.C.; Morrison, B.B., "Social Networking," *Computer* , vol.41, no.2, pp.97,100, Feb. 2008
  doi: 10.1109/MC.2008.61
  Abstract: In the context of today's electronic media, social networking has come to mean individuals using the Internet and Web applications to communicate in previously impossible ways. This is largely the result of a culture-wide paradigm shift in the uses and possibilities of the Internet itself. The current Web is a much different entity than the Web of a decade ago. This new focus creates a riper breeding ground for social networking and collaboration. In an abstract sense, social networking is about everyone. The mass adoption of social-networking Websites points to an evolution in human social interaction.
  keywords: {Internet;Web sites;groupware;social sciences computing;Internet;Web applications;Web sites;electronic media;human social interaction;social networking;Collaboration;Context;Educational institutions;Explosions;Facebook;Humans;IP networks;Internet;Social network services;Web sites;Facebook;MySpace;Wikipedia;YouTube;how things work;social networking},
  URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4454412&isnumber=4454380

Step 1b Deconstructed

"In the context of today's electronic media, social networking has come to mean individuals using the Internet and Web applications to communicate in previously impossible ways."

>> Defines the term "social networking", explains terminology

- Annotation in RED
- Text and annotation, both go into corpus

Step 2

- Compile the expressions and sentences, clearly marking the reusable parts. The non–reusable parts are the gaps to be filled in. This procedure should be part of your learning life – never stop doing it.

Step 2 Deconstructed

- Reuse:
- “In the context of today's ___, __ has come to mean __ using the __ to communicate in previously impossible ways.”

- These added to corpus

Step 3

- Classify the materials according to the schemata of a scientific paper
- Can be done using two options
  - How?

Step 3, Option 1

- Assign the expressions to the pre-defined scheme for the various parts of an article, together with the selection (e.g. an expression taken from a component from the Introduction is automatically classified as such)
  - Advantage: easier and quicker
  - Disadvantage: user does not practice reshuffling the material

Step 3, Option 2

- Select a large number of expressions (hundreds!) and only classify them later
  - Advantage: more efficient to learn how to reuse the expressions
  - Disadvantage: more time-consuming

Step 3 Deconstructed

- This text can be <context>
- “In the context of today's ___, __ has come to mean __ using the __ to communicate in previously impossible ways.”

- These added to corpus

Step 4

- Practice filling in the gaps with your own material and/or based on other examples

Step 4 Deconstructed

- “In the context of today's *technology*, *app* has come to mean *a computer application* using *a mobile platform* to __ in previously impossible ways.”

- These added to corpus

Step 5

- Start playing with the pieces, identifying different combinations that appear in the original texts and creating your own combinations ("the bricks are the same but the houses will be different").
- In this process, try to enrich the possibilities by selecting other expressions (Step 2)
- Keep practicing filling in the gaps (Step 4)

Step 5 Deconstructed

- “In today's *technology*, *app* has come to mean *a computer application that enables users to communicate…”*

- These added to corpus

Step 6

- Start all over again with the selected expressions, now classifying them according to rhetorical messages (e.g. *describe, contrast, confirm, define, compare, introduce*)
- The idea is to have a collection of expressions to be retrieved as you wish to state specific contents
- Keep selecting further expressions and filling in the gaps

Step 6 Deconstructed

- "This is largely the result of a culture-wide paradigm shift in the uses and possibilities of the Internet itself."

>>Describes a situation/current state

- These added to corpus

Step 7

- Start working with full text passages, rather than only with separate sentences
- Repeat the procedures of combining pieces, as in Step 5.
- Now is the time to learn using connectives efficiently. Compile a list of expressions including *however, in contrast, indeed, on the other hand, furthermore, nevertheless, since, because*, etc.

Step 7 Deconstructed

- Group sentences into paragraphs
- Compile expressions

- These added to corpus

Step 8

- Produce full section of paper
  - select the subcomponents
  - implement them by reusing material from your earlier practices
  - fill in the gaps – help may be obtained by retrieving material from the practices.
  - check the use of connectives and the text coherence

Step 8 Deconstructed

- ….

- These added to corpus

Step 9

- Check the section for typos and other surface errors
- Eliminate unnecessary words
- Check the consistency of the subcomponents and their inter-relationship.
- Analyze the contents for completeness and accuracy

Ethel Schuster, Ph.D.

Step 9 Deconstructed

- EDIT your text
- These added to corpus

Format?

- Can create
  - Multiple files
    - One text file with corpus for each selected paper

  OR

  - One text file with corpus for ALL papers

End ?

- ONE single file, two versions
    1. electronic AND
    2. Printed

Details (1)

- Can provide corpus in the format of the Chusaurus
- Include consolidated names for components and strategies
  for each part of the paper

Details (2)

- Essential that students
  - classify and annotate sentences
    - >>what is/is not "reusable”
- Must show that
  - understood how to prepare corpus
  - went through process of classifying contents
  - Internalize knowledge of corpus

How much?

- How big should the corpus be?
- Is there a minimum size?

Size of Corpus (1)

- Minimum size includes annotated sentences and material for all
  - components and
  - strategies

  for every section of a paper

Size of Corpus (2)

- Example: Slide 8
- Shows
  - all components and subcomponents of an Abstract
  - all the strategies (22)

>>Student must find and include sentences for ALL

Size of Corpus: Let's do it

- Include 3-6 sentences for each strategy, generates more than 100 sentences for Abstract
- Same applies for other sections

Size of Corpus: Let’s do it

Include

- lists of sentences with markers

  e.g., however, in addition, nevertheless, hence, thus, contrary to

- rhetorical strategies

  e.g., describe, contrast, exemplify, emphasize

Size of Corpus: DONE!

- When student does all these
  - final corpus will have
    - At least 30–40 pages of a Word document
    - 1 or 1.5 spacing, font 12

Automatic Annotation: Brat

- Web-based tool for text annotation
  - i. e., adding notes to existing text documents
- Designed for *structured* annotation
  - Notes have fixed form that can be automatically processed and interpreted by computer

Tutorial

- http://143.107.182.99/tutoriais/

Survey

http://www.surveymonkey.com/s/2L77ZRF