

# Testing GPT-4o for the semantic annotation of (competing) doublets

Cristina Lara-Clares  
University of Jaén

Cristina Fernández-Alcaina  
Charles University in Prague

*Biennial of Czech Linguistics  
Prague, September 19, 2024*



# Contents



1. STATE OF THE  
ART



2. METHOD



3. RESULTS AND  
DISCUSSION



4. CONCLUSIONS



5. FURTHER  
RESEARCH

# LLMs and automatic annotation



Large Language Models (LLMs): advanced machine learning models that use deep neural networks to process and learn data  
Significant improvements in automating a wide range of language tasks



Previous experiments with LLMs for the annotation of syntactic (e.g. Ambridge & Blything 2024), semantic (e.g. Gilardi et al. 2023; Frei & Cramer 2023), pragma-discursive (Imamovic et al. 2024; Yu et al. 2024) features



Key aspects:  
Crucial role of prompt engineering techniques  
Promising results → potential reduction of manual work

# Competing doublets: previous work



Competition (narrow definition)

Same base (e.g. *wrap*<sup>N</sup> – *wrapper*)  
Different affix (e.g. zero-affix and *-er*)  
Semantic overlap in certain contexts



Partial synonymy > sense-level analysis



Previous corpus-based work relied on manual annotation (Lara-Clares 2023; Fernández-Domínguez 2024)

# Competing doublets: limitations



Limitations related to sense level classification

esp. with polysemous words (Hilpert 2020: 9)

*interpretative indeterminacy* (Geeraerts 2010: 197): e.g. *Jean is a plain (=ugly/simple?) girl*



Methodological limitations

The ideal: team of human annotators, specialized skills, extensive training

The reality: time-consuming, lack of resources, human errors, inconsistencies



Q: can we use LLMs to reduce manual work and its methodological limitations?

# Method

1. Selection of doublets from previous research:
  - *wrap*<sup>N</sup> – *wrapper* (Lara-Clares 2023)
  - *black*<sup>V</sup> – *blacken* (Bauer et al. 2010; Fernández-Alcaina 2021)

Semantic overlap supported by lexicographic data, e.g.,

**wrap**  
— NOUN —

2.a. A loose garment or article of feminine dress used or designed to envelop or fold about the person; a shawl, scarf, or the like.

**wrapper**  
— NOUN —

1.2.b. A shawl, mantle, etc., for wearing about the person.

# Method

1. Selection of doublets from previous research:
  - *wrap*<sup>N</sup> – *wrapper* (Lara-Clares 2023)
  - *black*<sup>V</sup> – *blacken* (Bauer et al. 2010; Fernández-Alcaina 2021)

Semantic overlap supported by lexicographic data, e.g.,

**black**  
— VERB —

2.c. *transitive*. To apply black colour to (the body, esp. the face), esp. as a disguise or in order to play the role of a black person; = **blacken** v. 2b. Cf. *to black up* at Phrasal verbs.

**blacken**  
— VERB —

2.b. To apply black colour to (the body, esp. the face), esp. as an aid to concealment or in order to play the role of a black person. Cf. *to blacken up* at Phrasal verbs.

# Method

1. Selection of doublets from previous research
2. Classification and tagging of the senses of the derivatives: OED + Cambridge, Collins, Longman, Merriam-Webster and Wiktionary

<i>black</i> <sup>v</sup>	BECOME_BLACK	To be or become black
	MAKE_BLACK	To make black, blacken
	CLEAN_WITH_BLACKING	To clean and polish (shoes, a stove, etc.) with blacking
	PAINT_FACE_BLACK	To apply black colour to (the body, esp. the face), esp. as a disguise or in order to play the role of a black person
	DRAPE_WITH_BLACK	To drape with black (obsolete)
	BRUISE	To bruise or discolour the area surrounding the eye by a blow (esp. to black a person's eye)
	DEFAME	To stain, sully (a person's name); to defame, represent (a person) in bad light
	BLACKLIST	To declare to be black in an industrial dispute
	BEHAVE_AS_POACHERS	To operate as one of the poachers known as 'the Blacks' (obsolete)
	DRAW_IN_BLACK	To draw or shade in black
	BLACKMAIL	To blackmail (slang, now rare)



# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation (*British National Corpus*)



Derivative	N
<i>wrap</i> <sup>N</sup>	284
<i>wrapper</i>	178
<i>black</i> <sup>V</sup>	167
<i>blacken</i>	328
Total N = 957	

# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation (*British National Corpus*)



*Starting to feel a little frazzled round the edges, she went and had a bath, donned a nightdress and cotton **wrap**, and brushed out her hair. It was too early to go to bed yet - though she'd an idea that while this worry was on her mind she*

# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation



*Starting to feel a little frazzled round the edges, she went and had a bath, donned a nightdress and cotton **wrap** , and brushed out her hair. It was too early to go to bed yet - though she'd an idea that while this worry was on her mind she*

LOOSE\_CLOTHING: 'A loose garment that is worn tied around the body, especially the shoulders'

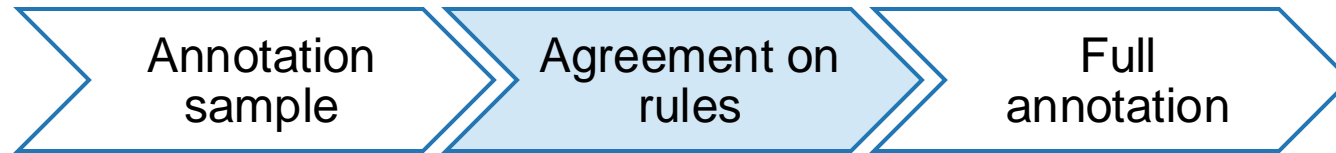
OUTER\_CLOTHING: 'An outer garment used to be worn as a defence against wind or bad weather'

...

NONE: If the noun 'wrap' does not reflect any of the senses from 1 to 14 above.

# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation



*Starting to feel a little frazzled round the edges, she went and had a bath, donned a nightdress and cotton **wrap** , and brushed out her hair. It was too early to go to bed yet - though she'd an idea that while this worry was on her mind she*

LOOSE\_CLOTHING: 'A loose garment that is worn tied around the body, especially the shoulders'

OUTER\_CLOTHING: 'An outer garment used to be worn as a defence against wind or bad weather'

NONE: If the noun 'wrap' does not reflect any of the senses from 1 to 14 above.

+

INDOOR\_GARMENT: 'A loose piece of clothing that is worn in the house, sometimes over other clothes'

# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation



*Starting to feel a little frazzled round the edges, she went and had a bath, donned a nightdress and cotton **wrap** , and brushed out her hair. It was too early to go to bed yet - though she'd an idea that while this worry was on her mind she*

INDOOR\_GARMENT: 'A loose piece of clothing that is worn in the house, sometimes over other clothes'

# Method: Problematic cases

*'I heard that,' she told him in the same language. He gave her a bold smile as she turned to get her 'wrap', then he and the groundsman lifted the holidaymaker and carried him across the grass to the medical centre.*

- LOOSE\_CLOTHING: 'A loose garment that is worn tied around the body, especially the shoulders'
- OUTER\_CLOTHING: 'An outer garment used to be worn as a defence against wind or bad weather'
- INDOOR\_GARMENT: 'A loose piece of clothing that is worn in the house, sometimes over other clothes'

# Method: Problematic cases

*'I heard that,' she told him in the same language. He gave her a bold smile as she turned to get her 'wrap', then he and the groundsman lifted the holidaymaker and carried him across the grass to the medical centre.*

- LOOSE\_CLOTHING: 'A loose garment that is worn tied around the body, especially the shoulders'
- OUTER\_CLOTHING: 'An outer garment used to be worn as a defence against wind or bad weather'
- INDOOR\_GARMENT: 'A loose piece of clothing that is worn in the house, sometimes over other clothes'
- UNSP\_CLOTHING: 'A piece of clothing which cannot be identified as either LOOSE\_CLOTHING, OUTER\_CLOTHING or INDOOR\_GARMENT from context'
- NONE: If the noun 'wrap' does not reflect any of the senses from 1 to 14 above.

# Problematic cases

*sooner you were done, you had to start again you understand? Mm Cos you had to get, do it with lights, and every window was '**blacked**' over, oh terrible. Terrible to work in it. But er, we, we got through it all, didn't we? We got through it.*

- MAKE\_BLACK: 'To make black'
- BO\_LIGHTS\_OFF: 'To extinguish or obscure lights; to cause a building, city or the like to become dark'
- BO\_OBLITERATE: 'To obliterate with black, especially as a form of censorship or protection'



# Problematic cases

*sooner you were done, you had to start again you understand? Mm Cos you had to get, do it with lights, and every window was '**blacked**' over, oh terrible. Terrible to work in it. But er, we, we got through it all, didn't we? We got through it.*

- MAKE\_BLACK: 'To make black'
- BO\_LIGHTS\_OFF: 'To extinguish or obscure lights; to cause a building, city or the like to become dark'
- BO\_OBLITERATE: 'To obliterate with black, especially as a form of censorship or protection'

# Problematic cases

*sooner you were done, you had to start again you understand? Mm Cos you had to get, do it with lights, and every window was '**blacked**' over, oh terrible. Terrible to work in it. But er, we, we got through it all, didn't we? We got through it.*

- **MAKE\_BLACK**: 'To make black'
- **BO\_LIGHTS\_OFF**: 'To extinguish or obscure lights; to cause a building, city or the like to become dark'
- **BO\_OBLITERATE**: 'To obliterate with black, especially as a form of censorship or protection'

# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation
4. Design and testing of the prompts
  - GPT4o (06/08/2024)
  - API

# Initial prompt: zero-shot CoT

*You are an expert linguist working on semantics and your aim is to check in which of the following sentences, the nouns 'wrap' or 'wraps' reflect one of the meanings that are provided below.*

*Specifically, the steps to follow are:*

- 1. Read the sentences containing the noun 'wrap' or 'wraps'.*
- 2. Annotate the sentences using the following labels (a-c):*
  - a) "MATERIAL" if the meaning expressed by the noun 'wrap' or 'wraps' refers to "material used to cover or protect objects";*
  - b) "CLOTHING" if the meaning expressed by the noun 'wrap' or 'wraps' refers to "an article of clothing that may be wrapped around a person";*
  - c) "NONE" if the noun 'wrap' or 'wraps' does not reflect any of the senses above or if the word "wrap" is a verb.*
- 3. Justify your answer.*

*The sentences you have to annotate are the following (numbered from 1 to 284) and with the noun 'wrap' or 'wraps' marked by simple quotation marks:*

- 1. The following assembler language "wrap around" provides the minimum acceptable error handling capability. The "'wrap' around" recognises and acts on the following events: An escape condition (the <ESCAPE> key having been pressed whilst*

*[...]*

# Initial prompt: zero-shot CoT

*You are an expert linguist working on semantics and your aim is to check in which of the following sentences, the nouns 'wrap' or 'wraps' reflect one of the meanings that are provided below.*

*Specifically, the steps to follow are:*

- 1. Read the sentences containing the noun 'wrap' or 'wraps'.*
- 2. Annotate the sentences using the following labels (a-c):*
  - a) "MATERIAL" if the meaning expressed by the noun 'wrap' or 'wraps' refers to "material used to cover or protect objects";*
  - b) "CLOTHING" if the meaning expressed by the noun 'wrap' or 'wraps' refers to "an article of clothing that may be wrapped around a person";*
  - c) "NONE" if the noun 'wrap' or 'wraps' does not reflect any of the senses above or if the word "wrap" is a verb.*
- 3. Justify your answer.*

*The sentences you have to annotate are the following (numbered from 1 to 284) and with the noun 'wrap' or 'wraps' marked by simple quotation marks:*

- 1. The following assembler language "wrap around" provides the minimum acceptable error handling capability. The "'wrap' around" recognises and acts on the following events: An escape condition (the <ESCAPE> key having been pressed whilst [...]*

# Final prompt: few-shot CoT

*You are an expert linguist working on semantics. Your aim is to distinguish the different senses the verb 'to black' can express in different contexts. Your task consists of three steps:*

*Step 1: read the sentences (numbered from 1 to 167) and with the word 'black' marked by simple quotations marks:*

*1 that has an extremely good record on nursery education and I hope that no one will believe the attempt to try and 'black' the good record that the Conservative and that have the Conservatives had have in this area. The budgets of course*

*Step 2: learn the following definitions (1-15) for the verb 'to black'. Each definition has a label in capital letters and an example representing the meaning of the definition using XML tagging.*

*1. BE(COME)\_BLACK: 'To be or become black, or to take on a black colour'*

*EXAMPLE: 'As she leaned against the stone cross, the sky <BECOME\_BLACK>blackened</BECOME\_BLACK> with clouds.'*

*2. MAKE\_BLACK: 'To make black'*

*EXAMPLE: 'She had very large eyes and used to <MAKE\_BLACK>black</MAKE\_BLACK> her eyelashes.'*

*3. APPLY\_BLACKING: 'To clean and polish (shoes, a stove or the like) with blacking'*

*EXAMPLE: 'When he <APPLY\_BLACKING>blackened</APPLY\_BLACKING> the boots of the rich businessmen by day, they were friendly enough.'*

*[..]*

*15. NONE: If the verb 'black' does not reflect any of the senses from 1 to 14 above.*

*Step 3: Annotate each of the sentences provided (1-167) using the labels BE(COME)\_BLACK, MAKE\_BLACK, APPLY\_BLACK, PAINT\_FACE\_BLACK, BRUISE, DEFAME, BLACKLIST\_BOYCOTT, DRAW\_IN\_BLACK, BLACKMAIL, BO\_OBLITERATE, BO\_LIGHTS\_OFF, BO\_CONSCIOUSNESS, NOUN, ADJECTIVE, or NONE. Present the results following the format: "SX: LABEL".*

*Step 4: Justify your answer.*

# Final prompt: few-shot CoT

You are an expert linguist working on semantics. Your aim is to distinguish the different senses the verb 'to black' can express in different contexts. Your task consists of three steps:

**Step 1: read the sentences** (numbered from 1 to 167) and with the word 'black' marked by simple quotations marks:

1 that has an extremely good record on nursery education and I hope that no one will believe the attempt to try and 'black' the good record that the Conservative and that have the Conservatives had have in this area. The budgets of course

**Step 2: learn the following definitions** (1-15) for the verb 'to black'. Each definition has a label in capital letters and an example representing the meaning of the definition using **XML tagging**.

1. BE(COME)\_BLACK: 'To be or become black, or to take on a black colour'

**EXAMPLE:** 'As she leaned against the stone cross, the sky <BECOME\_BLACK>blackened</BECOME\_BLACK> with clouds.'

2. MAKE\_BLACK: 'To make black'

**EXAMPLE:** 'She had very large eyes and used to <MAKE\_BLACK>black</MAKE\_BLACK> her eyelashes.'

3. APPLY\_BLACKING: 'To clean and polish (shoes, a stove or the like) with blacking'

**EXAMPLE:** 'When he <APPLY\_BLACKING>blackened</APPLY\_BLACKING> the boots of the rich businessmen by day, they were friendly enough.'

[..]

15. NONE: If the verb 'black' does not reflect any of the senses from 1 to 14 above.

**Step 3: Annotate each of the sentences provided (1-167) using the labels** BE(COME)\_BLACK, MAKE\_BLACK, APPLY\_BLACK, PAINT\_FACE\_BLACK, BRUISE, DEFAME, BLACKLIST\_BOYCOTT, DRAW\_IN\_BLACK, BLACKMAIL, BO\_OBLITERATE, BO\_LIGHTS\_OFF, BO\_CONSCIOUSNESS, NOUN, ADJECTIVE, or NONE. Present the results following the format: "SX: LABEL".

**Step 4: Justify your answer.**

# Sample output

1: DEFAME  
The sentence refers to defaming or tarnishing the Conservative's good record, aligning with the sense of staining someone's reputation.

2: BO\_CONSCIOUSNESS  
The context involves a person experiencing a temporary loss of consciousness, which matches the definition of blacking out in this sense.

3: BO\_CONSCIOUSNESS  
This sentence describes an incident where someone lost consciousness, fitting the definition of a blackout.

4: BO\_CONSCIOUSNESS  
The phrase "blacking out" is used to describe a loss of consciousness, aligning with this particular sense.

5: BO\_CONSCIOUSNESS  
The context implies a loss of consciousness, fitting the definition of blacking out.

6: BRUISE  
The sentence describes a physical altercation where someone was hit in the eye, resulting in a bruise or discoloration.

→ *black<sup>V</sup>*

280: INDOOR\_GARMENT  
281: INDOOR\_GARMENT  
282: LOOSE\_CLOTHING  
283: LOOSE\_CLOTHING  
284: UNSP\_COVERING

## Step 4: Justification

- Material: Instances where 'wrap' refers to a protective layer of paper/plastic/metal (e.g., S1, S14, S101).
- Blanket: Used in contexts referring to warmth/comfort (e.g., S2, S267).
- UNSP\_COVERING: Where 'wrap' is a covering but not specified (e.g., S3, S284).
- Loose Clothing: Clothing loosely worn around the body (e.g., S4, S89).
- Outer Clothing: Refers to clothing meant for outdoor protection.
- Indoor Garment: Clothing worn inside, often over other clothes (e.g., S6, S116).

→ *wrap<sup>N</sup>*



# Method

1. Selection of doublets from previous research:
2. Classification and tagging of the senses of the derivatives
3. Human semantic annotation
4. Design and testing of the prompts
5. Analysis of the results

# Results

Krippendorff's alpha (1970, 2012) for interrater agreement

- Suitable for multiple labels
- Accounts for chance agreement
- Two types of results:
  - Human 1 – Human 2
  - (Super)Human – GPT-4o

$$\alpha = \frac{p_a - p_e}{1 - p_e}$$

Krippendorff (1970, 2012), Castro (2017)

# Results

	HUMAN 1 – HUMAN 2		HUMAN – GPT-4o	
<i>wrap</i> <sup>N</sup>	0.780	Fair agreement	0.699	Fair agreement
<i>wrapper</i>	0.723	Fair agreement	0.478	Low agreement
<i>black</i> <sup>V</sup>	0.755	Fair agreement	0.651	Low agreement
<i>blacken</i>	0.895	Excellent agreement	0.0165	Low agreement

Krippendorff's alpha (1970, 2012):  
1: perfect agreement/reliability  
1- 0.800: excellent agreement  
0.800 - 0.667: fair agreement  
0.0667 - 0: low agreement  
≤0: complete disagreement/  
no reliability

# Results

	HUMAN 1 – HUMAN 2		HUMAN – GPT-4o	
		Annotators!		
<i>wrap</i> <sup>N</sup>	0.780	Fair agreement	0.699	Fair agreement
<i>wrapper</i>	0.723	Fair agreement	0.478	Low agreement
<i>black</i> <sup>V</sup>	0.755	Fair agreement	0.651	Low agreement
<i>blacken</i>	0.895	Excellent agreement	0.0165	Low agreement

Krippendorff's alpha (1970, 2012):  
 1: perfect agreement/reliability  
 1- 0.800: excellent agreement  
 0.800 - 0.667: fair agreement  
 0.0667 - 0: low agreement  
 0 or lower: complete disagreement (or no reliability)

# Results

	HUMAN 1 – HUMAN 2		HUMAN – GPT-4o	
<i>wrap</i>	0.780	Fair agreement	0.699	Fair agreement
<i>wrapper</i>	0.723	Fair agreement	0.478	Low agreement
<i>black</i>	0.755	Fair agreement	0.651	Low agreement
<i>blacken</i>	0.895	Excellent agreement	0.0165	Low agreement

Krippendorff's alpha (1970, 2012):  
1: perfect agreement/reliability  
1- 0.800: excellent agreement  
0.800 - 0.667: fair agreement  
0.0667 - 0: low agreement  
0 or lower: complete disagreement (or no reliability)

# Results

	HUMAN 1 – HUMAN 2		HUMAN – GPT-4o	
<i>wrap</i>	0.780	Fair agreement	0.699	Fair agreement
<i>wrapper</i>	0.723	Fair agreement	0.478	Low agreement
<i>black</i>	0.755	Fair agreement	0.651	Low agreement
<i>blacken</i>	0.895	Excellent agreement	0.0165	Low agreement

Krippendorff's alpha (1970, 2012):  
1: perfect agreement/reliability  
1- 0.800: excellent agreement  
0.800 - 0.667: fair agreement  
0.0667 - 0: low agreement  
0 or lower: complete disagreement (or no reliability)

# Results

	HUMAN 1 – HUMAN 2		HUMAN – GPT-4o	
<i>wrap</i>	0.780	Fair agreement	0.699	Fair agreement
<i>wrapper</i>	0.723	Fair agreement	0.478	Low agreement
<i>black</i>	0.755	Fair agreement	0.651	Low agreement
<i>blacken</i>	0.895	Excellent agreement	0.0165	Low agreement

Krippendorff's alpha (1970, 2012):  
 1: perfect agreement/reliability  
 1- 0.800: excellent agreement  
 0.800 - 0.667: fair agreement  
 0.0667 - 0: low agreement  
 0 or lower: complete disagreement (or no reliability)

'ADJECTIVE' vs other labels

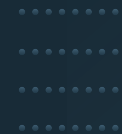
# Conclusions

- (Very) early stage but promising results
  - Prompts plays a crucial role and needs careful (re)design
  - Dissimilar results, performance close to human annotators (*wrap/black* vs *wrapper/blacken*) → data require closer inspection
- Limitations
  - Does not allow replicability – constant updates + *generative* language + need of API (cost; not accessible to everyone)
  - Prompting engineering still being tested and developed + learning curve
  - Unclear cross-linguistic validity; performance may depend on the availability of LLMs trained with data from other languages
  - (Still) time-consuming
    - Detailed semantic description (inevitably incomplete due to lexicographic limitations), esp. if polysemy
    - Need for careful examination



# Further research

- Expand to full clusters (i.e. groups of competitors)
- Data from different corpora, both synchronic (e.g. COCA) and diachronic (e.g. EHBC) (BUT limitations related to copyright still unclear)
- Tests with other LLMs (e.g. Gemini, Llama, Claude) and/or even new versions of GPT (e.g. Open AI o1, “a reasoning model for hard problems [...] performs similar to PhD students”)



# Testing GPT-4o for the semantic annotation of (competing) doublets

Cristina Lara-Clares  
clclares@ujaen.es

Cristina Fernández-Alcaina  
alcaina@ufal.mff.cuni.cz

*Biennial of Czech Linguistics  
Prague, September 19, 2024*

