

# Languages & the Media 2021

13<sup>th</sup> International Conference and Exhibition on Language Transfer in Audiovisual Media

September 20 – 24, 2021, online

**Blanca Arias-Badia**

Universitat Pompeu Fabra

**Anna Matamala**

Universitat Autònoma de Barcelona

**September 23, 2021**

## **Audio Description from an E2U Language Perspective**

A corpus-based study

# Audio description and E2U language

- A proposal explored in Bernabé & Orero (2020).
- Discussed with AD and E2U language experts for the case of Catalan (Arias-Badia & Matamala 2020).
- Research conducted in the framework of the EASIT Project.



Co-funded by the  
Erasmus+ Programme  
of the European Union



# Theoretical proposal

(Bernabé & Orero, 2020)

## LEXICAL LEVEL

Consideration of word frequency, word length, and sense disambiguation.

- Avoiding words which include difficult syllables.
- Avoiding abstract words.
- Avoiding foreign words.
- Avoiding metaphors, irony, proverbs and idioms.

## MORPHOSYNTACTIC LEVEL

Consideration of word order and sentence length.

- Avoiding pronouns if reference is unclear.
- Avoiding passive and progressive constructions.
- Prioritising short sentences without coordinate or subordinate clauses.
- Avoiding verbal periphrases.

# Main conclusions drawn from the focus group with experts

(Arias-Badia & Matamala, 2020)

- E2U experts favour the proposal.
- AD experts are reluctant → they assume the main target of AD are persons with sight loss, but without cognitive disabilities. They believe providing 'self-evident' information may be perceived as patronising and could 'hinder their imaginative process.' → They propose an adapted audio introduction to cater for the needs of persons with cognitive disabilities.

# Our research question

Is current AD in Catalan  
already easy-to-understand?

*This study is understood as a step prior to  
validation with users.*

# Methodology

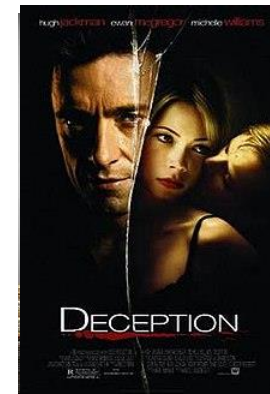
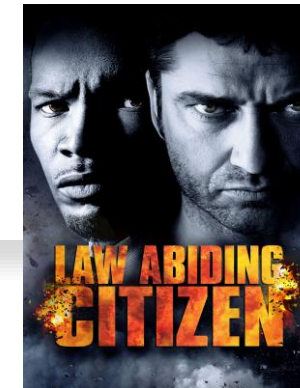
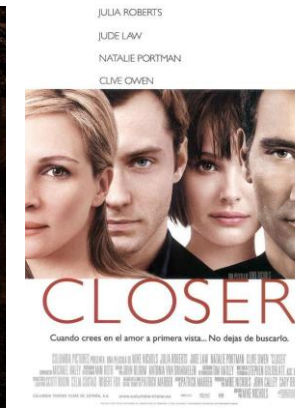
## MATERIALS

- ✓ AD of 9 films in Catalan (2004-2012): 46,908 tokens.
- ✓ E2R plot summaries of opera in Catalan as a term of comparison for one of the measures.

## TOOLS FOR SEMIAUTOMATIC ANALYSIS

- ✓ ContaWords.
- ✓ Desktop applications.

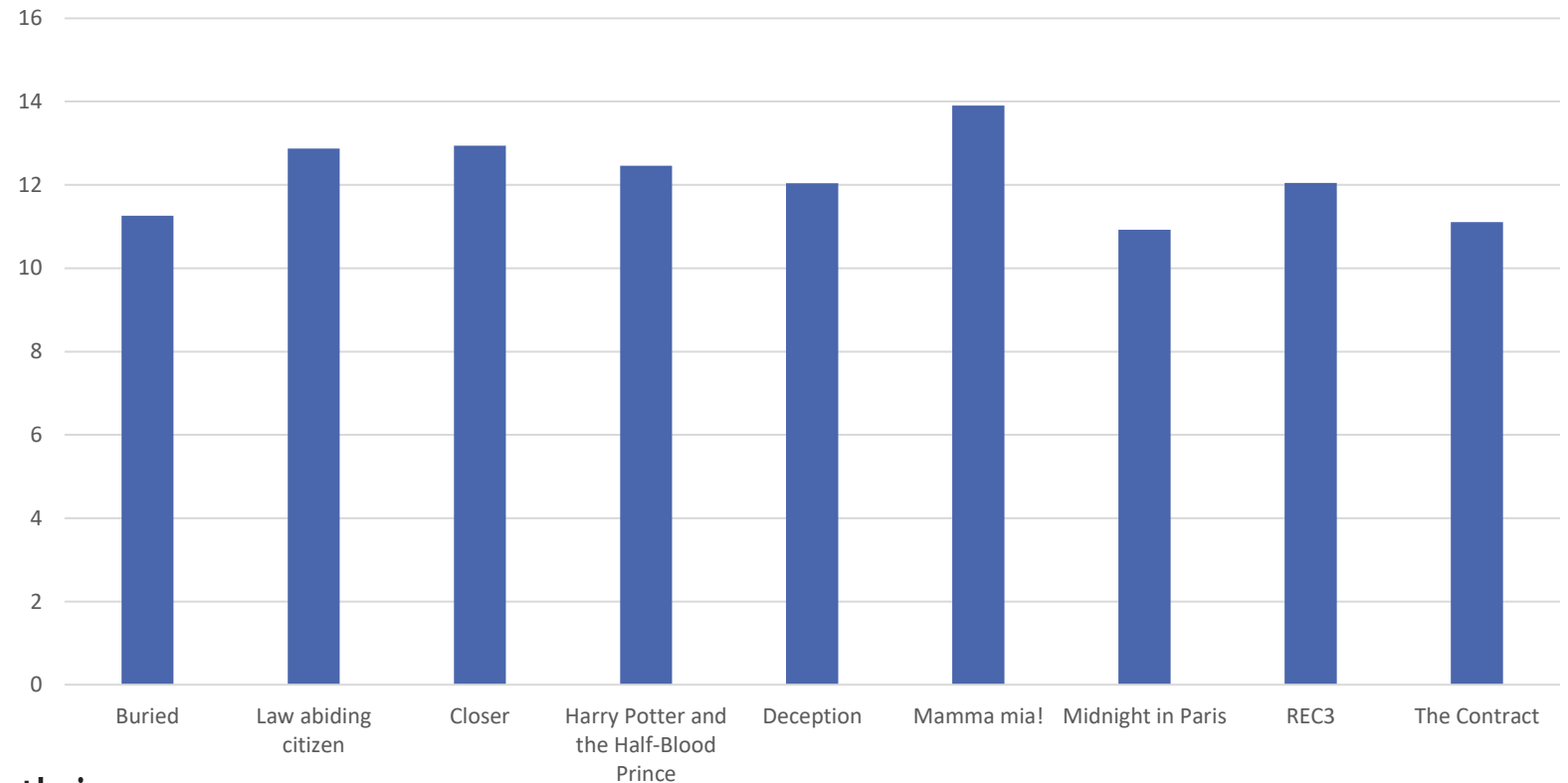
Complemented by manual annotation of a sample.



# Sentence complexity

# Sentence length

Average: 12.17



Average sentence length in...



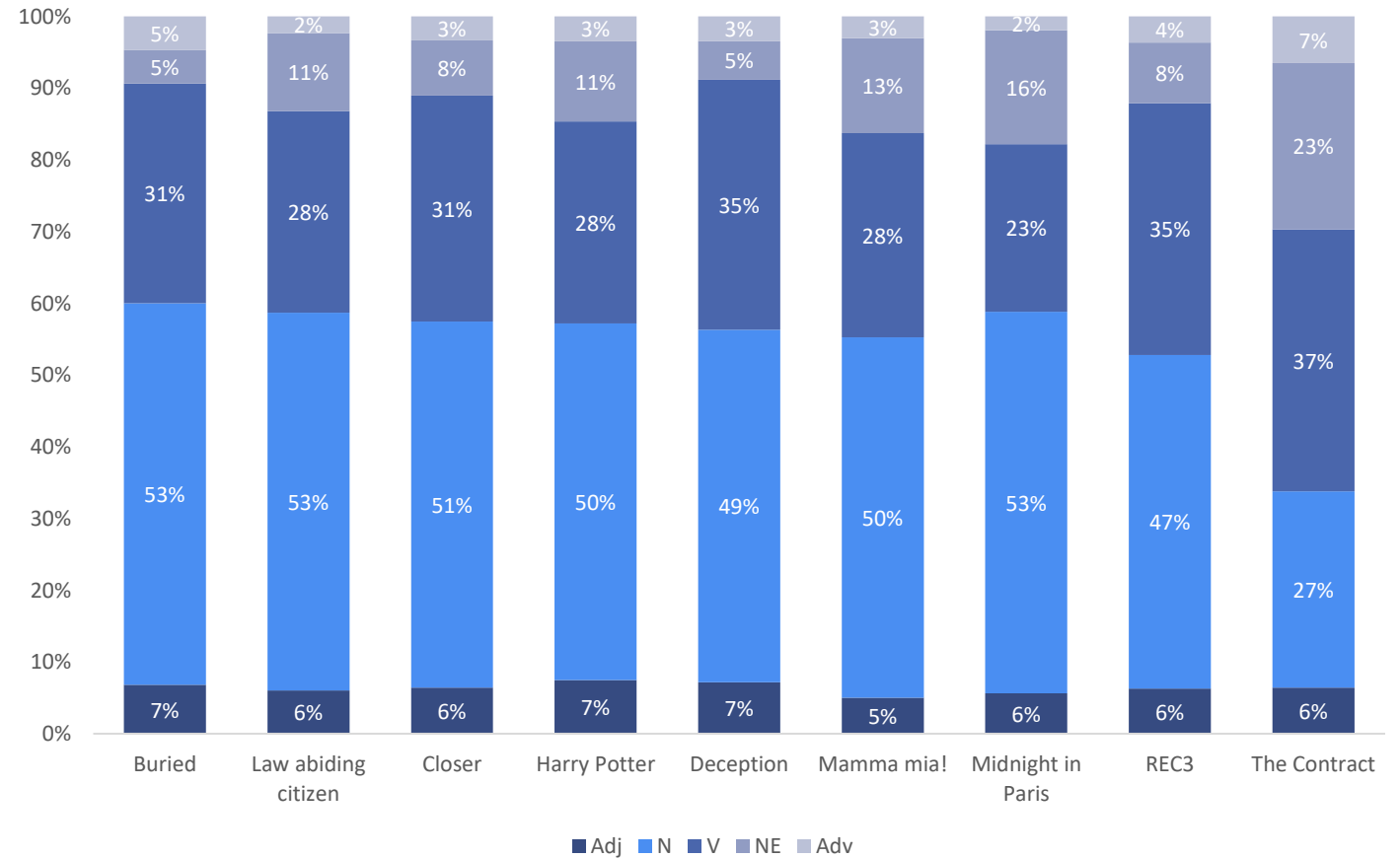
- general usage corpora in Catalan: 20.99;
- opera AD in Catalan: 13.71 (Hermosa-Ramírez, 2021);
- the VIW AD corpus in Catalan: 8.4 (Matamala, 2018).



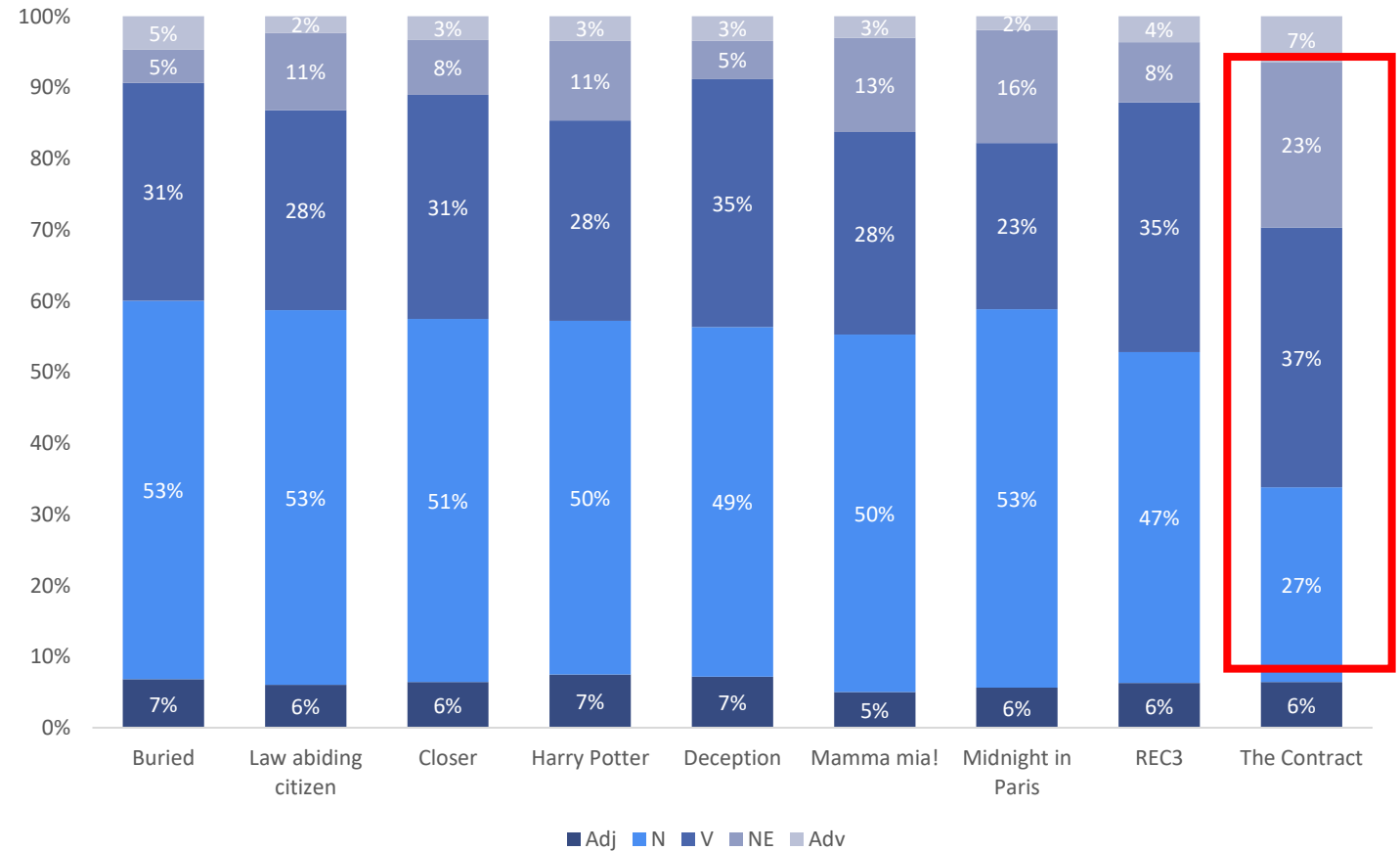
# Sentence length put in perspective

Corpus	Words/sentence	Source
<i>The Affair</i> -AD (ES)	9	Arias-Badia, 2021
VIW-professionals-AD (ES)	8	Matamala, 2018
VIW-professionals-AD (ES)	8	
Films and TV series-AD (NL)	14	Reviere, 2018
CoPP-television dialogue (EN)	8	Arias-Badia, 2017
CoPP-subtitling (ES)	7	
Journal for the general public (ES)	15	Mari Mut, 1998-2013
Scientific journal (ES)	25	
Legal documentation (ES)	55	

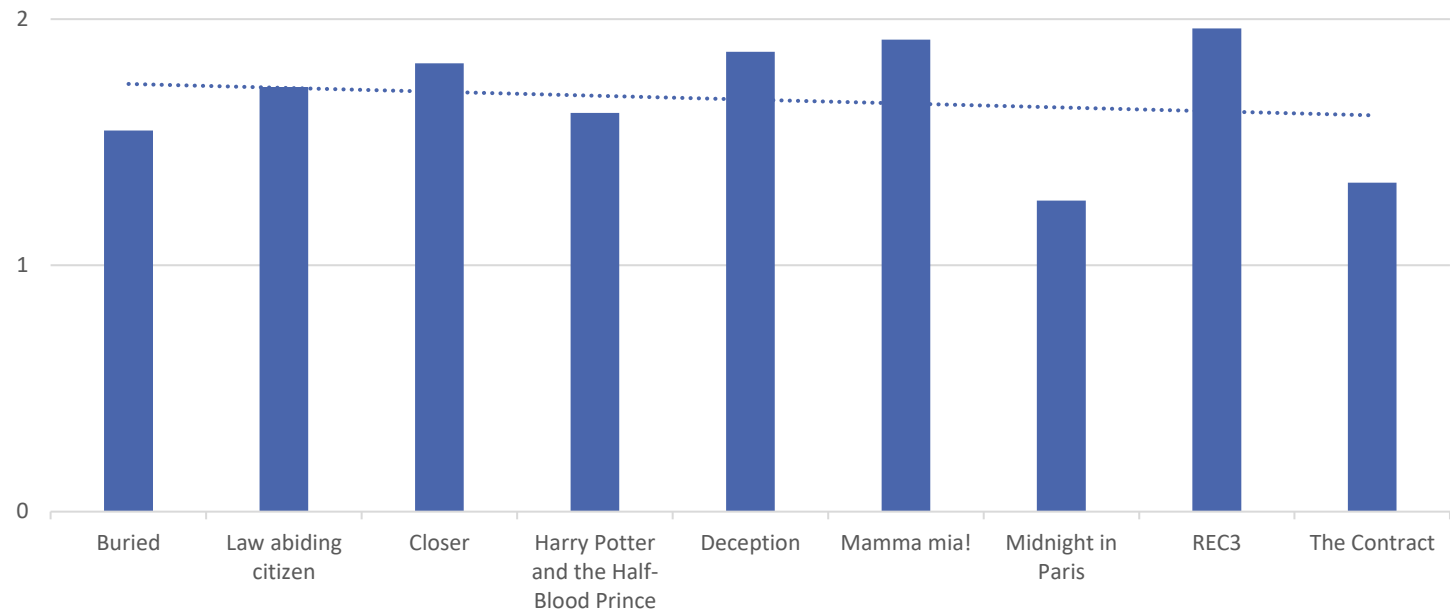
# PoS distribution



# PoS distribution



# Verbs per sentence



# Verbal periphrases in frequent bigrams

- *fa anar* (10)
- *deixa anar* (7)
- *deixa caure* (6)
- *va corrent* (4)
- *fa sortir* (4)
- *fa giravoltar* (4)
- *va mirant* (3)
- *fa passar* (3)
- *fa beure* (3)
- *entra corrent* (3)

# Verbal periphrases in frequent bigrams

- *fa anar* (10) 'uses'
- *deixa anar* (7) 'lets go'
- *deixa caure* (6) 'lets fall'
- *va corrent* (4) 'runs' PROGRESSIVE
- *fa sortir* (4) 'makes leave'
- *fa giravoltar* (4) 'makes turn'
- *va mirant* (3) 'looks' PROGRESSIVE
- *fa passar* (3) 'makes go through'
- *fa beure* (3) 'makes drink'
- *entra corrent* (3) 'comes in running' PROGRESSIVE

# Lexical analysis

# Aboutness

Words that typify the corpus

Three words present in the lists of the 30 most frequent words in all films:

- *mirar* 'look' – 53,198 oc. [CTILC]
- *mà* 'hand' – 51,954 oc. [CTILC]
- *fer* 'make/do' – 502,599 oc. [CTILC]



# Aboutness

is interesting from a  
narratological point of view

A by-product of analysing the kinds of information commonly provided by audio description is that **we also learn something about what events commonly happen in films**. Consider, for example, phrases that describe **characters looking at each other and at key objects**, phrases that indicate characters changing location and phrases that describe characters' expressions of emotions.

Salway (2007)

# Aboutness

Words that typify the corpus

**Names of characters:** *Jonathan, Paul, Harry, Sarah, Dunnigan...*

**Locations:** *carrer* ('street'), *terra* ('floor'), *bosc* ('forest'), *cuina* ('kitchen'), *presó* ('jail')...

**Objects:** *cotxe* ('car'), *taula* ('table')...

**Film-specific lexicon:** *vareta* ('wand'), *ensangonar* ('cover in blood'), *infectar* ('infect'), *encenedor* ('lighter').

*A note on SL/GL (Salway 2007).*

	Lexical density (TTR)	Vocabulary richness (lemmas/tokens)	Information load (lexical words/tokens)
Buried	0.24	0.18	0.45
Law abiding citizen	0.2	0.16	0.48
Closer	0.19	0.15	0.44
Harry Potter and the Half- Blood Prince	0.18	0.13	0.46
Deception	0.18	0.14	0.45
Mamma mia!	0.2	0.15	0.48
Midnight in Paris	0.3	0.25	0.49
REC3	0.19	0.14	0.46
The Contract	0.2	0.16	0.32

	Lexical density (TTR)	Vocabulary richness (lemmas/tokens)	Information load (lexical words/tokens)
Buried	0.24	0.18	0.45
Law abiding citizen	0.2	0.16	0.48
Closer	0.19	0.15	0.44
Harry Potter and the Half-Blood Prince	0.18	0.13	0.46
Deception	0.18	0.14	0.45
Mamma mia!	0.2	0.15	0.48
Midnight in Paris	0.3	0.25	0.49
REC3	0.19	0.14	0.46
The Contract	0.2	0.16	0.32

	Lexical density (TTR)	Vocabulary richness (lemmas/tokens)	Information load (lexical words/tokens)
Buried	0.24	0.18	0.45
Law abiding citizen	0.2	0.16	0.48
Closer	0.19	0.15	0.44
Harry Potter and the Half-Blood Prince	0.18	0.13	0.46
Deception	0.18	0.14	0.45
Mamma mia!	0.2	0.15	0.48
Midnight in Paris	0.3	0.25	0.49
REC3	0.19	0.14	0.46
The Contract	0.2	0.16	0.32
<b>Average</b>	<b>0.21</b>	<b>0.16</b>	<b>0.45</b>



Average lexical density (TTR) in...

- opera AD in Catalan: 0.39 (Hermosa-Ramírez, 2021).
- AD in other languages/products: 0.26 (EN, Arma); 0.51 (EN, Perego); 0.42 (Soler-Gallego); 0.38 (NL, Reviers).

# The Gunning Fog Index

A specific measure for E2U language

# Formula

Applied on samples of 100-150 words of each film

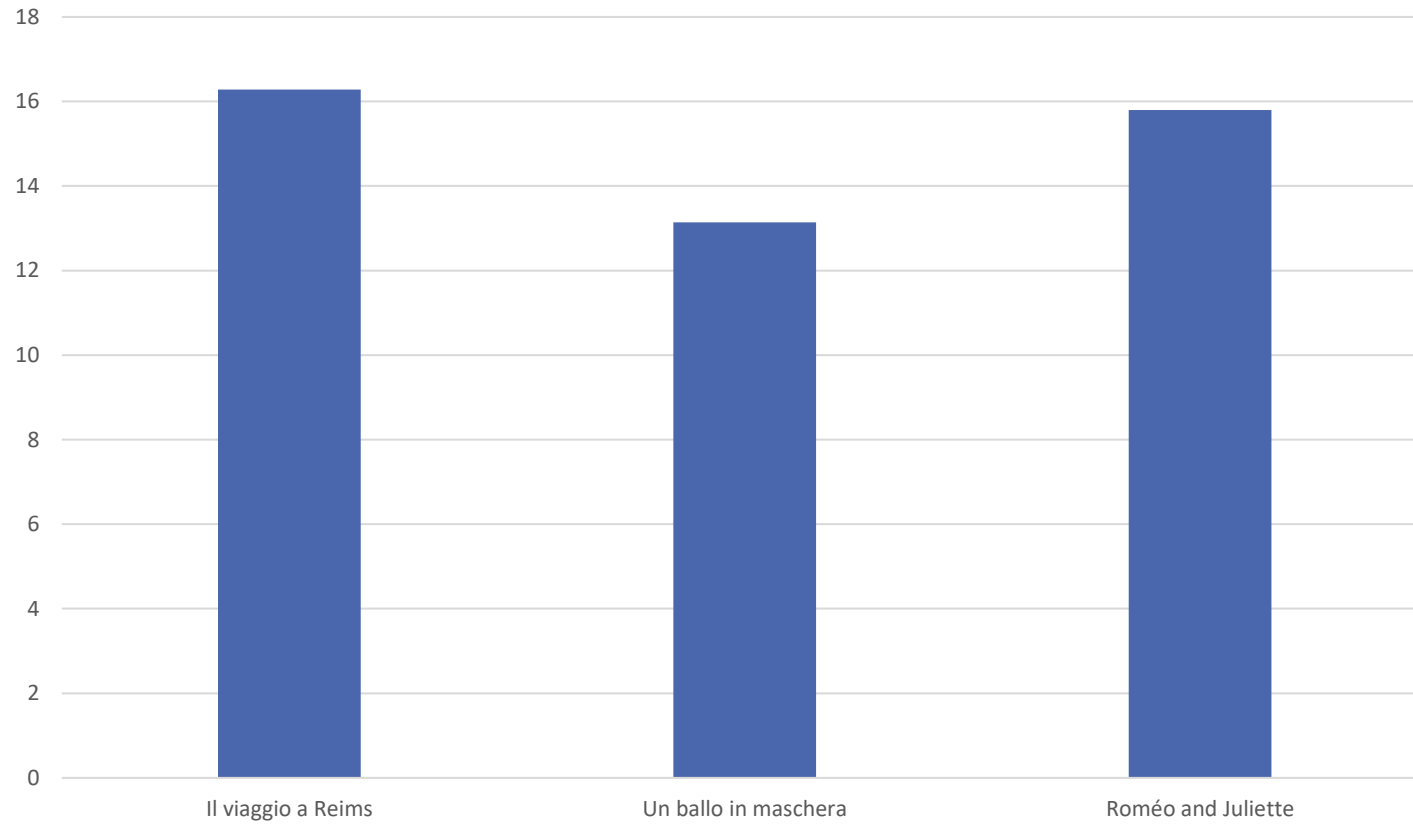
A= Average sentence length [Number of words/Number of sentences].

B= Percent Hard Words [“Complex” words/Number of words].

$$(A+B) * 0.4$$

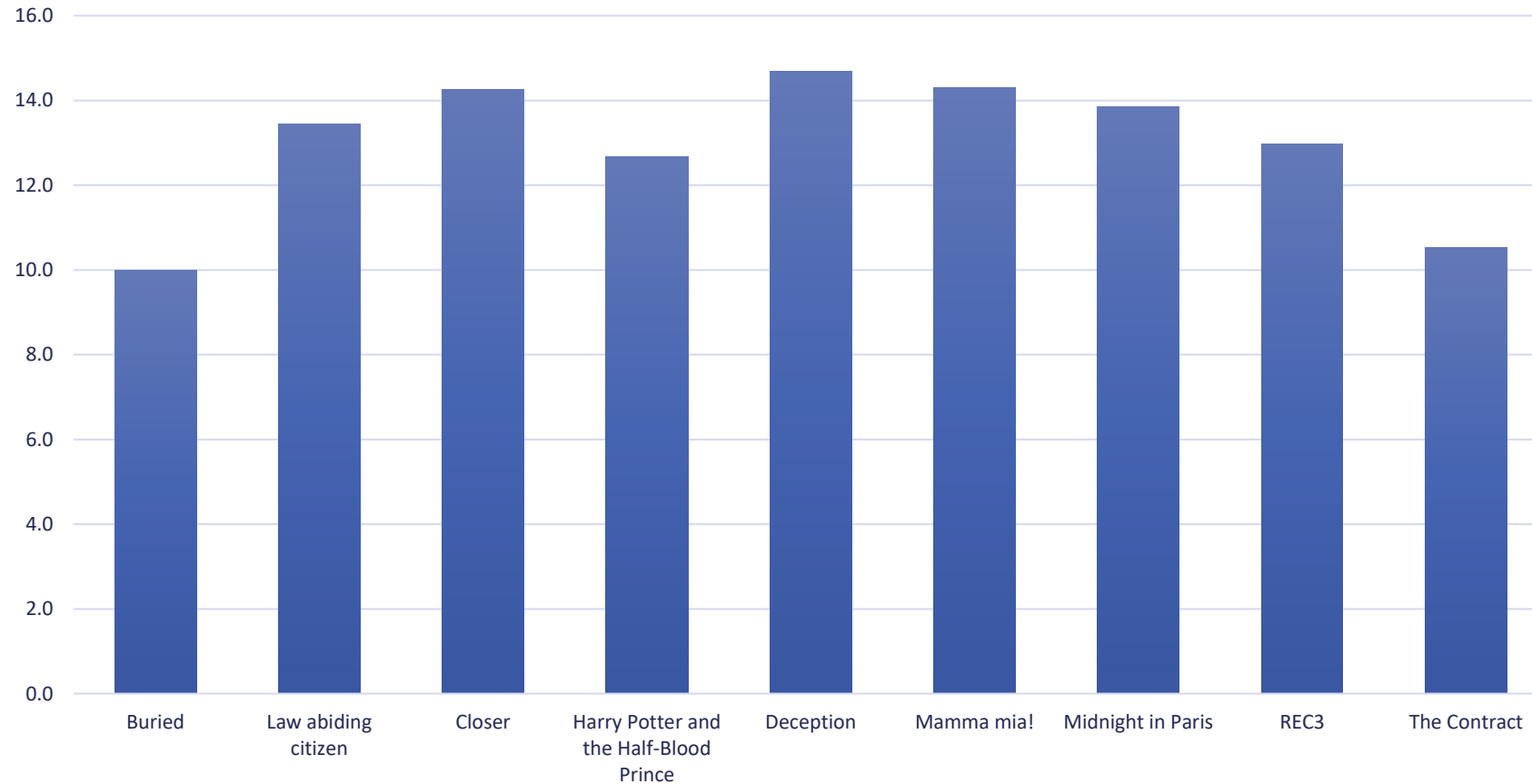


This measure has been used previously in the literature to compare the accessibility of texts written in the English language (Perego 2020).



The Gunning Fog Index of three E2R opera plot summaries validated by the Catalan E2R Association were taken as term of comparison.





None of the AD script excerpts in the subcorpus reaches a Gunning Fog Index of 15, as described for the validated plot summaries

# A closer look

at the potentially “most complex” AD  
in our corpus: *Midnight in Paris*

## STRUCTURE

118 AD units

145 simple sentences / 52 complex  
sentences, most of which are copulative  
sentences (44).

# A closer look

at the potentially “most complex” AD  
in our corpus: *Midnight in Paris*

## STYLE AND CONTENT

No literary tropes or figurative language.  
Only one lexicalised comparison: ‘En Gil se  
la queda mirant amb **uns ulls com  
taronges**’. – 81 collocates for *ull-taronja* in  
CTILC.

No neology or foreign words.

Almost all AD units use SVO in Catalan,  
which fosters clarity. 2 instances of non-  
neuter order.

# Final remarks

- A vast ground seems to be shared by professionals in AD and E2R in Catalan in terms of how texts are structured:
  - ✓ simple clauses are preferred;
  - ✓ neuter syntactic order is prioritised;
  - ✓ general usage lexicon is employed;
  - ✓ clarity in the identification of characters and objects is deemed a priority;
  - ✓ ...
- Now, these results are to be reinforced by testing them with users.

# References

- Arias-Badia, Blanca. 2020.** *Subtitling Television Series: A corpus-driven study of police procedurals*. Oxford: Peter Lang.
- Arias-Badia, Blanca. 2021.** The audio description of sex in *The Affair*. Advanced Research Seminar in Audio Description 2021. Universitat Autònoma de Barcelona. Barcelona (Spain), 27 January 2021.
- Bernabé, Rocío & Pilar Orero. 2020.** Easier audio description: A more readily understood accessibility service. In Sabine Braun and Kim Starr (eds). *Innovation in Audio Description Research*. London: Routledge. 55-75
- Arias-Badia, Blanca & Anna Matamala. 2020.** Audio description meets Easy-to-Read and Plain Language: results from a questionnaire and a focus group in Catalonia. *Zeitschrift für Katalanistik* 33/2020. 251-270.
- Hermosa-Ramírez, Irene. 2021.** The hierarchisation of operatic signs through the lens of audio description: A corpus study. *MonTI* 13. 184-219.
- Matamala, Anna. 2018.** One short film, different audio descriptions. Analysing the language of audio descriptions created by students and professionals. *Onomázein* 41. 185-207.
- Mari Mut, José A. 1998-2013.** *Manual de redacción científica*. Ediciones Digitales.
- Perego, Elisa. 2020.** *Accessible communication: A cross-country journey*. Berlin: Frank & Timme.
- Reviere, Nina. 2018.** *A corpus-based study into the linguistic features of a new, multimodal text type*. PhD thesis. University of Antwerp.
- Salway, Andrew. 2007.** A corpus-based analysis of audio description. In Jorge Díaz-Cintas, Pilar Orero and Aline Remael (eds). *Media for All. Subtitling for the Deaf, Audio Description, and Sign Language*. Amsterdam: Rodopi. 151-174.

## Authors' acknowledgements / Disclaimer

*Blanca Arias-Badia*

InfoLex is a research group funded by Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, under the SGR funding scheme (ref. code 2017SGR01366). This paper has been partially funded by the Department of Translation and Language Sciences at Universitat Pompeu Fabra.

*Anna Matamala*

TransMedia Catalonia is a research group funded by Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, under the SGR funding scheme (ref. code 2017SGR113).

The project EASIT has received funding from the Erasmus + Programme of the EU under the call Strategic partnerships for higher education, grant agreement No 2018-1-ES01-KA203-05275. The European Commission support for the production of this presentation 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.

TransMedia Catalonia and the funding organisms referred to in this presentation cannot be held responsible for any use which may be made of the information contained therein.

# Languages & the Media 2021

13<sup>th</sup> International Conference and Exhibition on Language Transfer in Audiovisual Media

September 20 – 24, 2021, online

**Blanca Arias-Badia**

Universitat Pompeu Fabra

**Anna Matamala**

Universitat Autònoma de Barcelona

**September 23, 2021**

## **Audio Description from an E2U Language Perspective**

A corpus-based study