Audio description from an easy-to-understand language perspective: A corpus-based study in Catalan

Blanca Arias-Badia, Universitat Pompeu Fabra, Barcelona
Anna Matamala, Universitat Autònoma de Barcelona

ABSTRACT

Providing accessible audiovisual content which caters for diverse user needs is one of the challenges of today's digitised society. Audio description (AD) has been recently proposed to adopt the principles of easy-to-understand language (E2U) to reach larger audiences (Bernabé-Caro and Orero 2020). A focus group developed in the context of the EASIT project (Arias-Badia and Matamala 2020) showed that some professionals were reluctant to accept easy AD, arguing that AD addresses persons who cannot see, rather than persons who cannot understand. Comments were also made indicating that current audio descriptions may already be easy to understand. Given the lack of research in this area, this article presents results of a corpus study conducted on nine AD scripts in Catalan and provides data on the extent to which current audio descriptions are already "easy" taking into account existing parameters linked to easy-to-understand language principles. The features under analysis are part of speech distribution, sentence complexity, the use of (in)frequent lexicon, word length, and lexical variation. The results show that audio description scripts in Catalan do hold features typically attributed to easy-to-understand language.

KEYWORDS

Audio description, easy-to-understand language, easy-to-read language, understandability, media accessibility, corpus-based studies, Catalan language.

1. Introduction: Background and aims of the study

Audio description (AD) translates visual elements into words (Snyder 2014; Remael *et al.* 2015; Fryer 2016). Audio description is an intersemiotic transfer mode in which the visual content, together with audio elements that may be difficult to understand without access to the visuals, is translated into spoken words (Maszerowska *et al.* 2014). These spoken words are generally inserted in the silent gaps of a wide array of recorded dynamic audiovisual productions such as films, cartoons, or documentaries. Audio description can also be used for live content (for instance, a conference, a theatre play) and static content (for instance, artworks) (Matamala and Orero 2016). Research on AD has focused on what to translate but also on how to translate it. In other words, it has dealt with content selection but also with language choice. AD has traditionally been addressed to those who cannot access the visuals, especially blind persons and persons with low vision, but research has also shown the benefits of AD for other target groups such as language learners, who benefit from watching the audiovisual content while listening to the audio description (Walczak 2016; Navarrete 2018).

Easy-to-understand (henceforth, E2U) language, in turn, is an umbrella term used to refer to different simplified language varieties (Maaß 2020; Perego 2020: 17) which range from easy language (also called easy-to-read) to plain language. Easy language is the most simplified form, with specific linguistic and formal features, and it is said to address persons with reading difficulties. Plain Language, on the other hand, falls at the other end of the spectrum and is expected to address lay citizens. Other terms are also being used such as "simple" or "clear language," showing the terminological fuzziness in the field.

In the context of the EASIT project (Matamala *et al.* 2021; Matamala 2022), and following the path of Bernabé-Caro and Orero's (2019, 2020) investigations, a question was asked: could E2U language principles be applied to AD? A focus group developed in the context of the EASIT project (Arias-Badia and Matamala 2020) showed that some professionals were reluctant to welcoming this option, arguing that AD addresses persons who cannot see, rather than persons who cannot understand. Comments were also made indicating that current audio descriptions are already easy to understand. Given the lack of research in this area, we aimed at taking a first step by asking the following research question: are existing audio descriptions easy to understand? Focusing only on AD in Catalan, this article aims to analyse to what extent current audio descriptions are already "easy" considering existing parameters linked to E2U language principles. The analysis uses a corpus of nine audio described films in Catalan and focuses only on aspects related to the written text. Analysing the voicing and delivery features and performing a user evaluation are fundamental, but outside the scope of this article.

The article begins with a revision of the main features of the language of AD according to the most relevant literature and with a description of the main characteristics of E2U language. It then presents the methodology, the corpus used, and the features selected for the corpus-based analysis. A discussion of the main results is presented next, before reaching conclusions and opening the door to future research.

2. The language of audio description (AD)

In the ADLAB guidelines, Taylor (2015: 46) discusses wording and style AD. He defines wording as "the ability to choose the right words in the right places," whereas style "is the result of the word choice of authors, along with their choice of sentence structure and appropriate use of figurative and idiomatic language." Taylor considers that AD "requires attention to both wording and style," which will be determined by a) time constraints, and b) the oral nature of the AD, which will be spoken and listened to. Taylor provides some general principles. In terms of lexical items, the author indicates the following:

> • Clear language and concrete vocabulary, unencumbered by jargon, unnecessary pomp and obscure vocabulary help with information processing and visualisation.
> • Precision and detail can be expressed by the use of colourful adjectives and adverbs or adverbial phrases [...].
> • A vivid language engages the listener and can be expressed, for instance, in verb variation [...].
> • The visual nature of a film can be reflected in the use of verbs of movement and simile, metaphor or other figures of speech [...] (Taylor 2015: 46).

The reference to a "clear language and concrete vocabulary" seems well aligned with easy-to-understand language, whereas "verb variation" and the use of figures of speech may clash with existing recommendations promoting the repetition of the same lexical unit to refer to the same element to avoid confusions.

In terms of grammar, the recommendations provided by Taylor (2015: 46) include favouring present tense and third person pronouns, using short sentences and avoiding subordination, following unmarked syntactic order (SVO in English), using simple phrases, and considering the spatio-temporal configuration of the visuals. Apart from these general principles, each source text will require specific choices, depending on the genre, time, place, filmmaker, and target audience.

Perego (2019: 119) also deals with the language of AD and gathers the many adjectives that have been used to define it: meticulous, meaning it "provides detailed, accurate and precise descriptions through well-chosen, clear (vs. obscure, jargon-rich) vocabulary"; visually intense, referring to "the depth and the force with which AD conveys visual details in words"; concise, due to the restricted time available, and usable, meaning "ADs that are easy to access and understand. In AD, usability is typically achieved through the use of plain syntax favouring short sentences and uncluttered constructions, as well as a logical organization of information" (2019: 120), a definition which also seems to be in line with easy-to-understand language principles, as will be shown in the next section.

Snyder (2014: 41) identifies four fundamentals of audio description: observation, editing, language, and vocal skills. Similarly, Fryer (2016: 58-65) devotes specific sections to word choice and creative use of language in her handbook. The standard ISO/IECTS 20071-21 *Information technology – User interface component accessibility. Part 21: Guidance on audio description* (ISO 2015: 11) indicates that audio describers "should present their information in a manner that can be easily understood by their intended users," making reference both to proper articulation and lexical choice. The standard acknowledges that different AD styles can be adopted, ranging from a newsreader style to a commentator style, first person or third person. There is a specific section which provides guidance on parts of speech, which recommends using "descriptive verbs" to "reduce repetition of common verbs" and "enhance audience experience and understanding" (ISO 2015: 23). In this regard, Fryer favours understanding

for the targeted end users, who in the context of her research are blind persons and persons with low vision, not necessarily persons with comprehension difficulties.

The Spanish standard UNE 153020 *Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías* (AENOR 2005) states that the style should be fluid, simple (*sencillo*), with direct sentences, avoiding cacophonies, redundancies, and poor idiomatic expressions. It also recommends using specific adjectives, rather than vague ones.

As far as Catalan is concerned, Puigdomènech *et al.* (2007) provided some recommendations for a future AD protocol in Catalan, based on the existing literature, in the framework of a Batista i Roca project. The study considers that the oral standard should be the basis for AD and that AD language should be adequate, understandable, and credible. In this regard, frequent expressions should be prioritised over archaic or technical ones. They also indicate that colloquialisms or dialectalisms should not generally be included. They provide specific advice on the use of articles and verbal tenses, indicating the present tense should be generally used. In terms of syntax, sentences should be short and easy to understand, avoiding information overload. The standard SVO order should generally be followed, and vocabulary should be clear, concise and at the same time rich. Bassols and Santamaria (2009) also make various linguistic proposals for Catalan AD: they recommend short and simple sentences, with some coordination, including one idea or a maximum of two per sentence. In terms of vocabulary, they are similarly in favour of specific and short words rather than long and abstract vocabulary. They do not see repetition as a problem of AD, as the AD units are interwoven with the dialogues, but suggest alternatives such as subject ellipsis or syntactic or semantic anaphora, among other linguistic devices. Apart from clarity, they advocate for concision through different linguistic mechanisms and a special attention to sentence order.

The recommendations summarised so far seem to indicate some shared ground with guidelines on easy-to-understand language, as will be shown in Section 2: references to "simple" or "clear" vocabulary which can be easily understand abound, next to references to a simple syntax. A key aspect is that the language of AD should be "understandable," as already mentioned by some of the authors above, but the target audience they generally have in mind are blind and partially sighted persons, not necessarily persons with comprehension difficulties. On the other hand, recommendations seem to favour a vivid and rich language, which cater for the AD target user needs but may imply a higher lexical complexity for certain audiences.

Corpus-based studies on the language of AD are not extensive: Salway (2007) analysed the specific features of a corpus of 91 scripts in British English in the context of the TIWO (Television Into Words) project. This corpus was also used by Arma (2011) to research the usage of adjectives. The TRACCE corpus, including 300 ADs in Spanish, has also been analysed by adopting a multimodal approach to audio description scripts (Jiménez Hurtado and Seibel 2012). In Dutch, Reviers *et al.* (2015) has used corpus linguistics tools to analyse a corpus of 17 scripts. The linguistic features of the Visuals into Words (VIW) corpus, including audio descriptions in Catalan, English and Spanish for the same short film (47 in total), have been analysed (Matamala 2018, 2019). All these studies have shed some light on the linguistic features of audio description in different languages but none of them have addressed a comparison with the principles of easy-to-understand language. Still, some of these studies have provided insights into the degree of difficulty of AD language as they have addressed elements which can be linked to it, such as part-of-speech distribution or sentence complexity. Reference to the results obtained by these studies will be made during our discussion. Finally, research on the machine translation of audio description has shown that although the language of AD seemed fit for machine translation due to its "idiosyncratic language" and its "relatively short and simple sentences," there are more challenges than expected, pointing at a higher complexity (Vercauteren *et al.* 2021: 245).

3. Easy-to-understand (E2U) language

Many terms are used to refer to E2U language in its different forms: easy-to-read, easy read, easy reading, easy language, plain language, simple language, simplified language, citizen language, and clear writing are some examples. In this article, E2U language is considered an umbrella term that refers to different simplified language varieties which range from easy-to-read (also referred to as easy language) to plain language (Maaß *et al.* 2021: 194).

Easy-to-read, according to ISO/IEC 23859-1 (ISO 2023: 2) is a language variety "in which a set of recommendations regarding wording, structure, design and evaluation are applied to make information accessible to persons with reading comprehension difficulties for any reason." More recently, easy-to-read has been referred to as "easy language" to account for the fact that it can be used both in written and spoken language.

Plain language is, according to the US Plain Writing Act of 2010, writing that is "clear, concise, well-organized, and follows other best practices appropriate to the subject or field and intended audience." Plain language started to fully develop in the 1960s in the USA and in the 1970s in the UK in citizen and legal communication (Mazur 2000; Montolío and Tascón 2020: Chapter 1), where sometimes the terms "citizen communication/language" or "clear communication" are used (Montolío and Tascón 2020).

Lindholm and Vanhatalo (2021: 18) describe the differences between easy language and plain language in three areas: types of documents, target groups, and degree of simplification:

> Whereas Plain Language is related to institutional documents, and aims to simplify legal language for non-professionals, the notion of Easy Language refers to making various texts or speech accessible to people who have difficulties reading and understanding standard language. As a language form, Easy Language is usually more simplified than Plain Language (2021: 18).

In terms of characteristics, both easy language and plain language share many characteristics, but easy language often shows a higher degree of simplification as well as specific layout features such as accompanying images to enhance comprehensibility.

Some of the main guidelines for easy language are: International Federation of Library Associations and Institutions (IFLA 2010), first published in 1997; International League of Societies for the Mentally Handicapped (ILSMH-EA, 1998), and Inclusion Europe (2009). There is also the international standard ISO/IEC 23859-1 (ISO 2023), already mentioned, and the Spanish national standard UNE153101EX (2018). Cutts (2020) is a key reference for plain language in English, next to the US Federal Plain Language Guidelines (PLAIN 2011).

The ISO standard, which aims to adopt an overarching approach including different degrees of simplification, highlights the need to use a vocabulary suited for the intended audience and acknowledges that some words (vague, abstract, non-frequent, etc.) may be more difficult to understand than others, a recommendation also included in the Spanish standard. Both standards recommend avoiding unnecessarily long sentences, generally aiming at including one idea per sentence. An easy-to-follow structure next to an appropriate style considering the content and the audience are favoured, and various recommendations on form and layout are included next to some guidance on the inclusion of paratextual elements to support the comprehension process (images, glossaries, etc.).

4. Methodology

To answer our research question, i.e., whether present AD in Catalan follows the principles of E2U language, two sets of materials were analysed: film AD scripts and opera plot summaries. The motivation for choosing this material was twofold. First, film AD is the most widespread type of AD—since this is, to the best of our knowledge, the first study to look at AD from the perspective of E2U language, we deemed it interesting to prioritise film AD, and the AD of blockbusters, to gather conclusions that may be applicable beyond our corpus of study. Second, as will be explained, the opera plot summaries used were validated as easy-to-read in Catalan. Since

AD synthetises the visuals of plot development, we considered these narrative and synthetic texts to be comparable for our research interests. In what follows, each of these materials are presented, and the features under analysis in each of them are reported, as well as the tools employed.

The main analysis reported here was done on a corpus of nine AD scripts in Catalan of films produced between 2004 and 2012. Permissions from the AD scripts authors for academic use of this material were gathered to conduct the study. Table 1 shows the basic data of these films as provided by FilmAffinity (https://www.filmaffinity.com), as well as the number of tokens in each AD. The total amount of tokens of this film AD corpus is 46,908.

Original title	**Date**	**Country**	**Director**	**Production company**	**Duration**	**Number of tokens in AD script**
Buried	2010	France, Spain, US	Rodrigo Cortés	Versus entertainment	93 min	2,714
Closer	2004	UK, US	Mike Nichols	Columbia Pictures	104 min	4,322
Deception	2008	US	Marcel Langenegger	20th Century Fox, Seed Productions, Rifkin-Eberts, Media Rights Capital (MRC)	108 min.	6,202
Harry Potter and the Half-blood Prince	2009	UK	David Yates	Warner Bros., Heyday Films	153 min.	7,601
Law Abiding Citizen	2009	US	F. Gary Gray	The Film Department, G-BASE, Warp Films, Evil Twins	108 min.	5,611
Mamma Mia!	2008	UK	Phyllida Lloyd	Universal Pictures, Littlestar Productions, Playtone	108 min	6,911
Midnight in Paris	2011	France, Spain, US	Woody Allen	Gravier Production	96 min	2,326

				s, Mediapro, Pontchartrain Productions, Televisión de Galicia (TVG), Versátil Cinema		
[●REC]³: Génesis	2012	Spain	Paco Plaza	Castelao Pictures, Canal+ España, Filmax, ICIC, Ono	81 min.	6,469
The Contract	2006	US	Bruce Beresford	Millennium Films	92 min.	4,752

Table 1. Basic data about the films included in the corpus of study.

The second type of materials used were opera plot summaries written in easy-to-read language in Catalan, which were validated by the Catalan easy-to-read association, Associació Lectura Fàcil. We selected three plots randomly from the list available at the Liceu Opera Barcelona website (<https://www.liceubarcelona.cat/ca/lectura-facil>). The three plot summaries under analysis belong to the three plays presented during the 2017-2018 opera season at the Liceu, namely *Il viaggio a Reims* (Gioachino Rossini), *Un ballo in maschera* (Giuseppe Verdi), and *Roméo et Juliette* (Charles Gounod). These texts served as a term of comparison to inform the results found in the main corpus. These texts include 2,476 tokens in total — the corpus size thus is a limitation of the present study.

The main corpus was analysed from a morphosyntactic and a lexical point of view, by considering features typically attributed to E2U language and by adopting a mixed methods approach. At the morphosyntactic level of language, the analysis yielded results on occurrence and distribution of different parts of speech (PoS), with a focus on lexical words—namely adjectives, adverbs, named entities, nouns, and verbs—and syntax complexity. The open access tool Contawords©, developed by the Institute for Applied Linguistics at the Universitat Pompeu Fabra (Barcelona, Spain) (http://contawords.iula.upf.edu), was used to run an automatic lemmatisation of the corpus and PoS tagging, as well as to obtain the most frequent bigrams in each AD script. As regards sentence complexity, the following aspects were considered: number and type of sentences, sentence length, occurrence of verbal periphrases, and verbs per sentence.

At the lexical level of language, the following features were analysed: corpus aboutness, lexical density, vocabulary richness, and information load. Corpus aboutness can be defined as the lexical types that typify a corpus as a whole (Oakes 2012). In order to gather this information, the 30

most frequent lexical words from each film were retrieved from the outputs provided by Contawords©. In relation to E2U language, our aim in scrutinising corpus aboutness was to determine whether infrequent words in Catalan, such as specialised terms or expressive neologisms, had a salient occurrence in the corpus. At the same time, corpus aboutness reveals which common words are repeated across different films in AD scripts in Catalan.

In order to study lexical variation in the corpus, lexical density was computed using the type/token ratio formula (TTR) and vocabulary richness was obtained by applying the formula number of lemmas/number of tokens. TTR has been reported to be an adequate measure for lexical variation in texts of up to 5,000 words (Baker 2006: 52) and has been previously used in AD corpus studies (Arma 2011, Perego 2019). To foster comparability of the results obtained, the WordSmith Tools© software was used to automatically retrieve the standardised type/token ratio (STTR) of each AD script. The information load of the corpus was computed using the formula number of lexical words/tokens. Following Halliday's (1985) and Biber *et al.*'s (1999: 55) definitions of lexical words, the following PoS categories were computed as lexical words in the analysis: adjectives, adverbs, nouns, and verbs. These tasks were also facilitated by the use of Contawords©, which allows the computation of homonymic lemmas as different PoS (e.g., *mira* may be the third person singular of the verb *mirar* ('to look') or a noun in Catalan). This tool lists named entities (i.e., proper nouns) separately from common nouns; since the mention of characters' names is relevant for the study of AD scripts, we decided to keep this differentiation in our presentation of results.

The last measure applied to the corpus involved both a morphosyntactic and a lexical approach to the data. We ran tests to obtain the Gunning Fog Index of the film AD corpus. In Perego (2020), this index is used to compare the difficulty entailed in understanding film AD and art AD in English. This index has been traditionally used as a readability formula that reveals the relation between syntactic complexity as determined by sentence length and the occurrence of complex words. Following the recommendations of the site ReadabilityFormulas (https://readabilityformulas.com), we took a random sample of 100–150 words of each film (henceforth, the subcorpus) and annotated them manually. Afterwards, we applied the following formula: average sentence length (obtained by dividing the number of words by the number of sentences) was added to the percentage of complex words (obtained by dividing them by the total number of words). The sum of these elements was then multiplied by 0.4. This formula is reproduced in Figure 1. In line with the E2U language principles, this formula understands complex words to be those with three or more syllables that are not proper nouns, combinations of easy words or hyphenated words, or non-personal forms of the verb (gerunds, participles, infinitives) (ReadabilityFormulas n.d.).

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

B= Percent Hard Words [("Complex" words/Number of words) * 100].

(A+B) * 0.4

Figure 1. Gunning Fog Index formula. Source: ReadabilityFormulas (n. d.).

The subcorpus was also used to explore word length in AD scripts. We ran automatic syllable counts of these excerpts with the open access tool offered by Softcatalà© (https://www.softcatala.org/sillabes). Since AD is prepared in written format but delivered via the aural channel, we deemed it pertinent to consider both the calculation of syllables obtained after applying graphic criteria and after applying phonetic criteria of the Catalan language. Both options are available in the tool used for the analysis. When graphic criteria are applied, the tool renders the syllable count by considering how a given word is written. For example: *por*|*ta-li-ho* ('take it to them') is a 4-syllable string if graphic criteria are applied. When phonetic criteria are applied, the tool considers elisions which are prototypical of spoken discourse. For example, in the previous word, *por*|*ta-li‿-ho*, an elision is typically found between the last two graphic syllables. Thus, the word results in a 3-syllable string if phonetic criteria are applied.

Taylor (2015: 46) explains that, "if time permits, more variation in sentence structure can be pleasant and engaging" in an AD script. In order to illustrate the contexts in which audio describers seem to opt for more complex language, the above morphosyntactic and lexical analyses were complemented by the manual annotation of one of the scripts showing the "most difficult" AD in accordance with the results obtained, considering lexical density, vocabulary richness, information load, and the Gunning Fog Index measure (from the film *Midnight in Paris*). In this manual annotation, the features under analysis were the occurrence of hypotactic and paratactic structures, i.e., subordination and coordination, sentence order, and the use of literary tropes or figurative language, as well as new and foreign words.

As has been mentioned above, opera plot summaries were used as a term of comparison. Namely, we contrasted sentence length as well as their Gunning Fog Index. Since these texts in Catalan have been validated by the main easy-to-read organisation in Catalonia as texts that are easy to understand, checking that their Gunning Fog Index is similar to or higher than the one in the AD corpus would confirm the suitability of the formula to account for texts in Catalan.

5. Morphosyntactic analysis: Results and discussion

This section presents the results yielded from the analysis of part-of-speech distribution in the AD corpus and of linguistic features associated with sentence complexity, namely sentence length, verbs per sentence, occurrence of hypotactic and paratactic structures, presence of verbal periphrases and deviation from unmarked sentence order. The results are contrasted with previous accounts relevant to the research.

5.1. Part-of-speech (PoS) distribution

The distribution of lexical PoS is shown in Figure 2. As can be seen, the occurrence of common nouns stands out. Their mean presence is of 48% of all the lexical words in the corpus. Only one of the films, namely *The Contract*, departs from this trend and shows a higher occurrence of verbs (37%) than common nouns (27%). It must be noted, however, that this script has the highest occurrence of named entities, i.e., proper nouns, which balances the lower occurrence of common nouns. The results are in line with previous research conducted on AD corpora in different languages, in which nouns were also the most frequent PoS (Reviers 2018; Matamala 2018; Hermosa-Ramírez 2021).

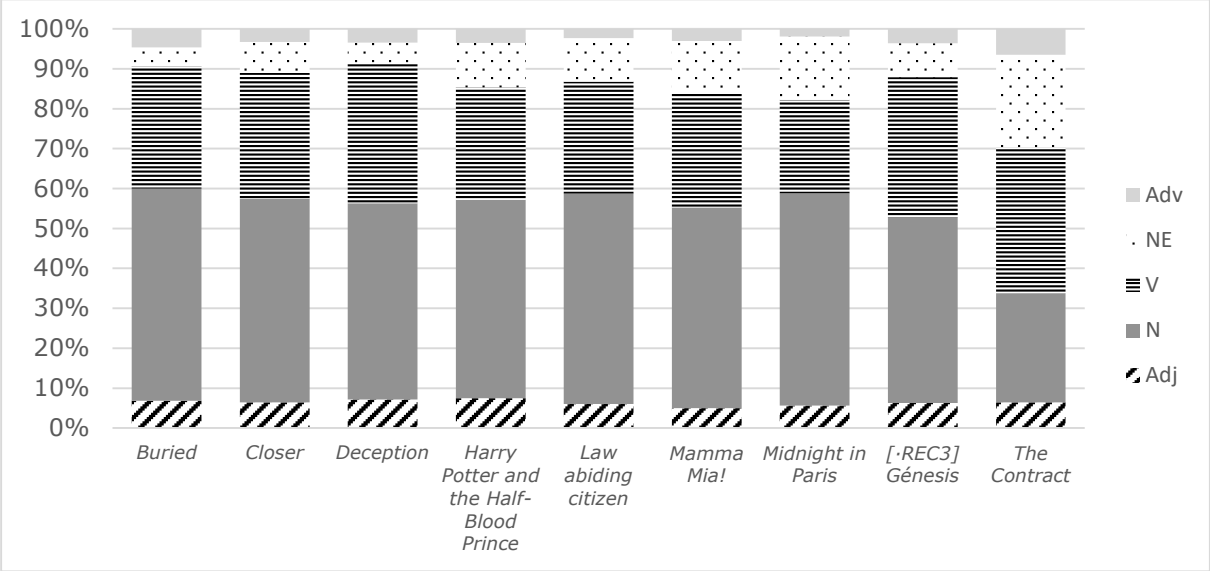

Figure 2. Distribution of lexical word categories in the corpus: adjectives, adverbs, named entities, nouns, and verbs.

From the point of view of E2U language, the European Commission (2012) recommends the use of verbal forms for the writing of clearer texts. According to their guidelines, "verbs are more direct and less abstract than nouns" (2012: 8). In this sense, the higher occurrence of nouns in the AD corpus shows room for improvement to achieve easier-to-understand ADs in Catalan.

The manual syntactic annotation of *Midnight in Paris* is useful to further reflect on this result. The analysis shows that only 29 sentences out of 197 in the AD script (14.7%) lack an explicit verb. Typically, sentences without verbs include isolated time expressions such as *De matí* ('In the morning'). While this lies beyond the scope of the present study, it would be interesting to run a reception test with end users to check whether standard E2U principles would be preferable here, that is, would adding a verb *to be* (e.g. *És de matí* ('(This) is in the morning') significantly improve comprehension in the context of AD scripts? Our hypothesis (only based on introspection) is that the language used is sufficiently clear. This intuition is reinforced by the report of previous experiences in writing texts aimed at persons with cognitive disabilities: in the case of the dictionary definitions of *Diccionario Fácil* (Plena Inclusión Madrid 2023), a team of experts in the needs of persons with cognitive disabilities disregarded the use of full sentences (including verbs) for the dictionary definitions of common nouns, which are systematically validated by end users (García Muñoz 2019).

5.2. Sentence complexity

Sentence complexity was found to be low in our corpus of AD scripts in Catalan after considering the features under study: sentence length, verbs per sentence, occurrence of hypotactic and paratactic structures, presence of verbal periphrases and deviation from unmarked sentence order. The mean sentence length of all the films in the corpus is 12.17 words per sentence (wps), with results in each film ranging from 10.9 wps to 12.9 wps, as shown in Figure 3.

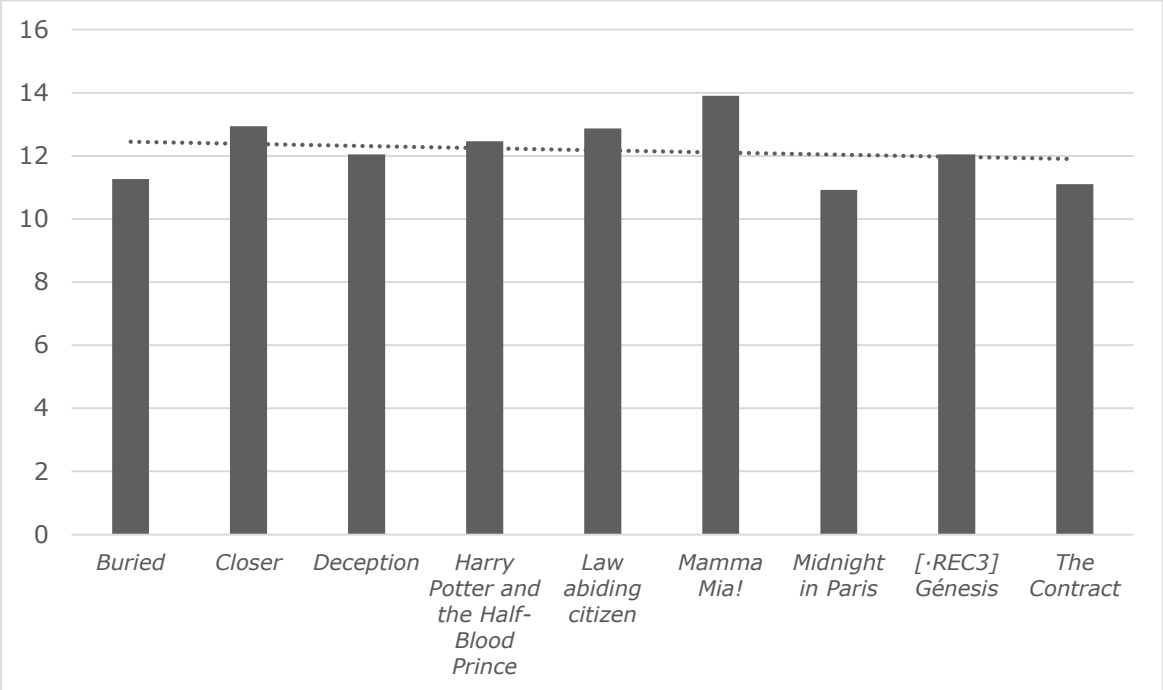

Figure 3. Mean sentence length in each AD script (words per sentence).

Inclusion Europe (2009: 11) recommends writers and adapters of easy-to-read and E2U content to "keep [their] sentences short." Likewise, the European Commission's (2012: 6) guidelines to foster clarity in writing establishes 20 words on average as a desirable standard for Spanish or French, as well as 25 words on average as a desirable standard for Italian; all three Romance languages close to Catalan. The sentences used in the AD of films in Catalan are shorter than the ones described in general usage corpora in Catalan (20.99 wps) as well as in opera AD in Catalan (13.71 wps) (Hermosa-Ramírez 2021). They are, however, longer than the sentences used in the Catalan AD of a short film in the VIW project (8.4 wps) (Matamala 2018). If we compare the results with corpus analysis of AD undertaken in other languages, sentences are also shorter than film and television AD in Dutch (14 wps) (Reviers 2018), but they are longer than the ones found in scenes with sexual content in the TV series *The Affair* in Spanish (9 wps) (Arias-Badia 2021). Finally, the mean sentence length in the easy-to-read plot summaries used as a term of comparison for this analysis scores higher: it ranges from 12.66 wps to 17.16 wps.

The mean verbs per sentence in the AD corpus was found to be 1.67, with scores ranging from 1.26 to 1.99 verbs per sentence, as shown in Figure 4. This means that most sentences are simple constructions, i.e., including one verb only, as is recommended for E2U content (Bernabé-Caro and Orero 2020).

The manual annotation of *Midnight in Paris* revealed 145 simple sentences versus 52 complex ones. Most complex sentences were found to be copulative sentences (44, i.e., 68.75%). In this sense, it must be noted that traditional research in the area of reading and listening comprehension established that coordinated structures were predicted to be easier to understand (Ardery 1980; Lust and Mervis 1980). It is also worth mentioning that studies in Psycholinguistics have thoroughly researched how long subordinate clauses including dependencies between non-adjacent words are more difficult to process in languages such as Spanish (López-Sancio 2022) or English (Grodner & Gibson 2005).

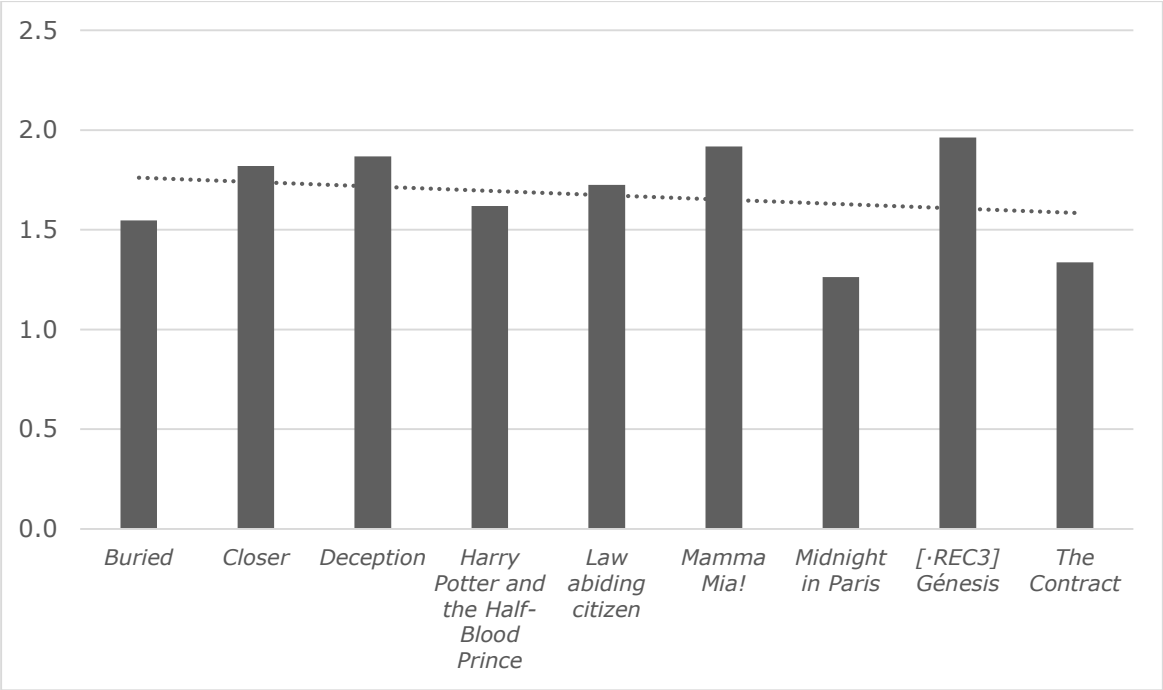

Figure 4. Mean verbs per sentence in each AD script.

It has been recommended to avoid verbal periphrases to favour comprehension (Bernabé-Caro and Orero 2020). The most frequent bigrams in each AD script were scrutinised as relevant data that reveal the occurrence of periphrases in the corpus. The ten periphrases displayed in Table 2 were identified in the analysis. As can be seen, they are common language use periphrases, such as colloquial forms to convey the meaning of 'use' (*fer anar*) or 'let go' (*deixar anar*). Three periphrases expressing progressive meaning were also identified (*va corrent*, *va mirant*, *entra corrent*). While Bernabé-Caro and Orero (2020) do not recommend the inclusion of progressive periphrases, it must be noted that their occurrence is low in the corpus (10 occurrences in total) and the instances found are very frequent periphrases in common language in Catalan.

Periphrasis	*English translation*	*Occurrence in the corpus*
fa anar	'uses'	10
deixa anar	'lets go'	7
deixa caure	'lets fall'	6
va corrent	'runs' (progressive)	4
fa sortir	'makes leave'	4
fa giravoltar	'makes turn'	4
va mirant	'looks' (progressive)	3

fa passar	'makes go through'	3
fa beure	'makes drink'	3
entra corrent	'comes in running' (progressive)	3

Table 2. Frequent periphrases in the corpus.

Sentence order is another aspect typically considered when promoting ways to express ideas clearly (European Commission 2012: 7). The recommendation is to follow a unmarked syntactic order (SVO) favouring the presentation of information in a structured manner. The manual analysis of *Midnight in Paris* showed a preference for unmarked order in the AD script. Most AD units (52 out of 118, that is, 44.1%) started with the subject of the action described. This is in line with the guidelines provided by PLAIN (n.d.: para. 2) for the Plain Writing Act of 2010, according to which a priority when producing E2U texts is to "make sure it's clear *who* does what." This type of structure was followed by AD units introduced by verbs, which occurred in sentences with implicit subjects (standard usage in Catalan), sentences with postponed subjects in verbs in which this is the standard structure (*arriba en Gil*, 'Gil arrives'), or impersonal sentences (*hi ha*, 'there is').

In some instances, the AD unit wants to highlight where or when the action takes places or, to a lesser extent, how the character or the filmic elements are presented. In this regard, 15 AD units (17.7%) start with an indication of where the action takes place: in six cases this is the only content of the AD unit (*En un restaurant luxós*, 'In a luxurious restaurant'), whereas in nine instances there is additional information (*Darrere de Nôtre Dame, la dona li llegeix el llibre*, 'Behind Nôtre Dame, the woman reads him the book') and the location is highlighted by moving it to the first position. In six AD units (7.08%) there is not a full sentence but just a noun or a noun phrase indicating what is seen: for example, *Instantànies de París* ('París snapshots'). Five AD units (5.0%) begin with expressions of manner, related to the attitude or position of the subject (*Estranyat*, 'Surprised') or the film technique used (*En primer pla*, 'In a close-up'). Four AD units start with an indication of the time (5.9%): *De nit* ('At night') (two occurrences), *De matí* ('In the morning') or *De dia* ('In the morning/In the afternoon/In the light of day'). Finally, there are also eight AD units (9.44%) which read the title or the cast, reproducing the written content on screen.

6. Lexical analysis: Results and discussion

This section presents the results of the study regarding the lexicon employed in film AD scripts in Catalan. The aspects considered are corpus aboutness, word length, and lexical variation.

6.1. Corpus aboutness: Use of frequent or infrequent lexicon

The 30 most frequent lexical words in each AD script, including named entities, were analysed to establish whether frequent or infrequent lexicon typifies AD scripts in Catalan. The appendix to this paper shows the words included in this part of the analysis, as well as their absolute and relative frequencies. As can be seen, the corpus is typified by a highly frequent lexicon in common language use, which speaks of its likely understandability. Note, for example, the four words that are to be found across all films: the verbs *fer* ('do/make') and *mirar* ('look'), and the nouns *cap* ('head') and *mà* ('hand'). Each of these words shows over 500,000 (*fer*) or over 50,000 (*mirar*, *cap*, *mà*) occurrences in the general use of language corpus of the Institute for Catalan Studies (CTILC).

The findings on corpus aboutness are in line with previous studies on frequent lexicon in AD scripts (Hermosa-Ramírez 2021; Matamala 2018; Reviers 2017, 2018; Salway 2007). The most frequent words refer to locations—*carrer* ('street'), *museu* ('museum'), *passadís* ('corridor')—, body parts—*mà* ('hand'), *cap* ('head'), *ull* ('eye')—, characters—proper names or pronouns, as well as nouns such as *noia* ('girl') or *home* ('man')—, verbs of movement —*passejar* ('stroll'), *anar* ('go'), *asseure's* ('sit')—or the expression of emotions—*somriure* ('smile'). Although objects do not stand out as especially frequent lexical types in the lists, nouns referring to objects typical of specific films also mark the lexicon of some AD scripts, such as *vareta* ('wand') in *Harry Potter and the Half-Blond Prince* or *encenedor* ('lighter') in *Buried*. Indeed, as noted by Salway (2007: 163), studying corpus aboutness in AD scripts is a means to better understand which elements are prioritised in filmmaking:

> 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: 163).

In the context of a study on AD, it is worth noting that the use of *mirar*, as well as of other frequent nouns connected to sight found in the analysis (*vista* ('look'), *mirada* ('gaze')) are salient in AD in different languages (Arias-Badia 2021; Hermosa-Ramírez 2021; Matamala 2018; Reviers 2018; Salway 2007). Again, the Catalan AD corpus is in line with these findings.

Manual annotation of *Midnight in Paris* also yielded relevant results in terms of word choice in AD scripts in relation to E2U guidelines. No instances of neology or foreign words were found. That suggests that AD scripts in Catalan prioritise the use of well-established lexicon in the language. Only one phrase used as a simile was found as a creative language device: *En Gil se la queda mirant amb uns ulls com taronges.* ('Gil stares at her with eyes like oranges', meaning that they are wide-open, showing

astonishment). Although this phrase is not included in relevant Catalan dictionaries such as the normative dictionary of the Institute for Catalan Studies (https://dlc.iec.cat/) or in the dictionary edited by Enciclopèdia Catalana (http://www.diccionari.cat/), it is lexicalised in common language use and there are over 80 occurrences of the lexical combinatorics of *ull* ('eye') and *taronja* ('orange') in the general language corpus CTILC (ctilc.iec.cat).

6.2. Word length

E2U language prioritises the use of short words. As explained in Section 4, one of the parameters taken into account in the Gunning Fog Index measure is the occurrence of words counting three or more syllables. Therefore, we computed the average number of syllables in the subcorpus. The results of this analysis are reproduced in Figure 5. As can be seen, none of the scripts reaches a mean of two syllables per word. The mean output is of 1.73 syllables per word by applying graphic criteria in the calculation, and of 1.56 syllables per word by applying phonetic criteria. This result shows a preference for short words in AD scripts in Catalan, thus favouring understanding, in accordance with E2U language principles.

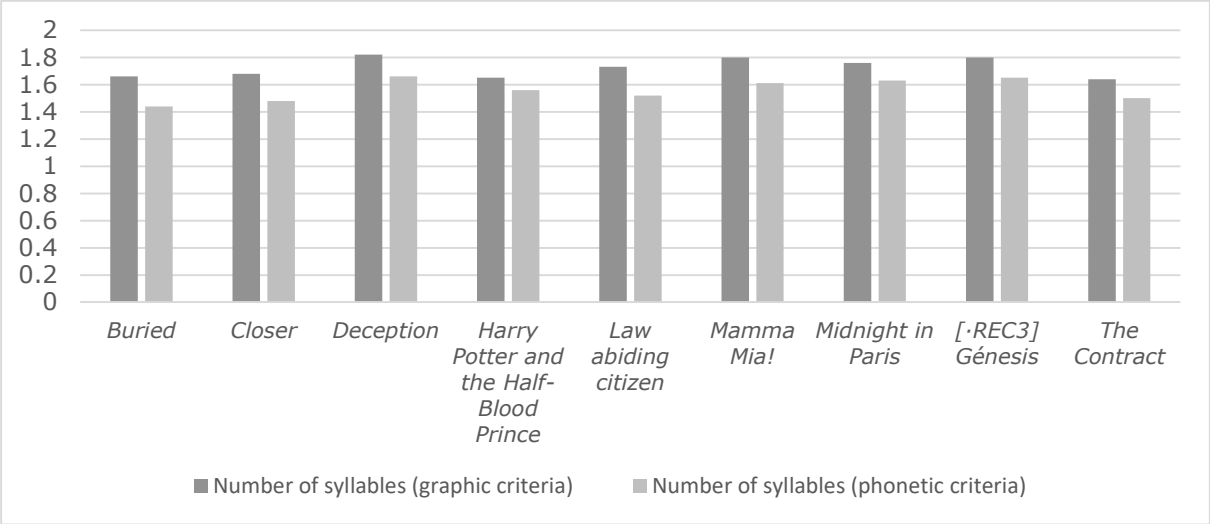

Figure 5. Mean number of syllables in the subcorpus.

6.3. Lexical variation

This section reports on results of three parameters that shed light on the lexical variety of the AD corpus in Catalan, namely lexical density (TTR), standardised type-token ratio (STTR), vocabulary richness, and information load. The results of applying the formulae described in Section 4 to obtain these measures is shown in Table 3.

	Lexical density (TTR)	***Standardised type-***	***Vocabulary richness***	***Information load (lexical***

		token ratio (STTR)	***(lemmas/tokens)***	***words/tokens)***
Buried	24%	38%	18%	45%
Law Abiding Citizen	20%	39%	16%	48%
Closer	19%	35%	15%	44%
Harry Potter and the Half-Blood Prince	18%	39%	13%	46%
Deception	18%	37%	14%	45%
Mamma Mia!	20%	38%	15%	48%
Midnight in Paris	30%	45%	25%	49%
REC3	19%	38%	14%	46%
The Contract	20%	37%	16%	32%

Table 3. Results of lexical variation measures applied to the corpus.

The average lexical density in the corpus is 21% (TTR) / 38% (STTR). As has been explained above, STTR is more easily comparable across texts of different lengths and in different languages (Baker 2006: 52). Thus, it is worth noting that this AD corpus follows the trend identified in previous research on AD, in different genres: Reviers (2018) reports a STTR of 38% in her study on film AD in Dutch; Soler Gallego (2018) finds a STTR of 42.5% in a study on museum AD in English; Hermosa-Ramírez (2022) reports a STTR of 35.6% for opera AD in Spanish and of 40.2% for opera AD in Catalan. The latter result signals that film AD shows less variation than opera AD when products in the same language, i.e., Catalan, are compared.

TTR results (21% in our corpus) need to be interpreted cautiously if compared with previous studies, since the measure is not directly comparable across languages and texts of various lengths. Hermosa-Ramírez (2021: 204) offers a synthesis of previous accounts of TTR in AD corpus research which is worth reproducing here for further information: "Arma's (2011) study on filmic AD reports 26.0% TTR for English and 31.5%

for Italian AD. On the other hand, Perego's (2019) study comprising 18 standalone ADs from the British Museum scores 51.07% TTR, a much higher ratio."

7. Application of the Gunning Fog Index measure

This section presents the results of applying the Gunning Fog Index (GFI) measure to the subcorpus. As mentioned in Section 4, this measure has been used previously in the literature to compare the accessibility of texts written in the English language (Perego 2020). We used the three opera plot summaries validated by the easy-to-read association in Catalonia, Associació Lectura Fàcil, as a term of comparison to check the GFI scores validated as accessible for the Catalan language. As shown in Figure 6, the GFI scores of the opera plots ranged from 13.1 (*Un ballo in maschera*) to 16.3 (*Il viaggio a Reims*).

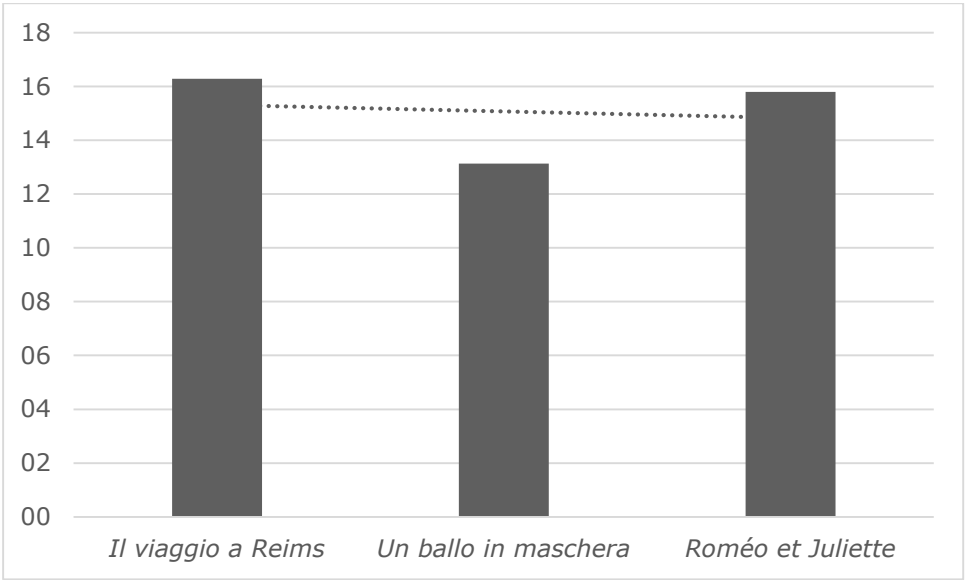

Figure 6. Gunning Fog Index of three opera plot summaries validated by the Catalan easy-to-read association, Associació Lectura Fàcil.

By taking these figures into account, it can be said that our film AD subcorpus scores well in terms of understandability. None of the AD script excerpts in the subcorpus reaches a Gunning Fog Index of 15—results range from 10 (*Buried*) to 14.9 (*Deception*) This means that their scores are, in average, below the ones for validated texts in easy-to-read language. See the results of the analysis in Figure 7.

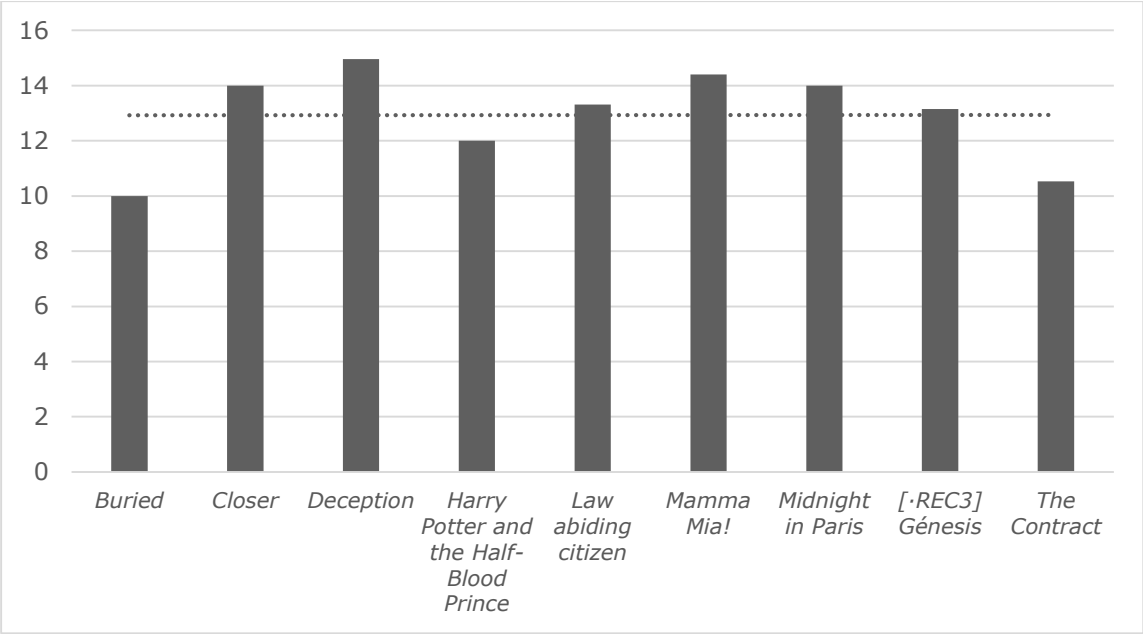

Figure 7. Gunning Fog Index of the subcorpus.

8. Conclusions and future work

This paper has approached film AD scripts in Catalan from the perspective of E2U language. A small corpus has been explored by a combination of qualitative and quantitative methods to address the extent to which film AD complies with the principles of E2U language. The results show that Catalan AD scripts share a large common ground with E2U principles: at the syntactic level of language, they prioritise the use of short, simple sentences, they avoid verbal periphrases, and tend to opt for the recommended structure to present information (i.e., presenting the subjects of actions first). At the lexical level of language, they prioritise short words, frequent lexicon in common language use, and show a lower lexical density than the one reported for AD in other languages and genres. Finally, they score well after the application of standard readability formulas such as the Gunning Fog Index.

The results speak in favour of fostering hybrid accessibility services like the one proposed by Bernabé-Caro and Orero (2020), namely E2U AD, in the sense that it seems that no major changes should be implemented in current AD practice in Catalan to be able to offer easy ADs as an output — perhaps a preference for verbal forms, rather than nouns, should be fostered in future AD scripts. This kind of service would be useful to cater for the needs of different user profiles. As we have argued in the Introduction, AD professionals were reluctant to accept this adaptation when asked about it *a priori* (Arias-Badia and Matamala 2020), but the results obtained in this study show that they may already be working in an E2U direction, perhaps inadvertently.

This study has limitations. As shown in previous methodological proposals (Biber 1995), small corpora—including up to 1,000 tokens, or 10 text samples—have proved useful to pinpoint the main features of specific registers or genres in exploratory studies even if they are not representative of a given population. However, larger corpora of both film AD and validated texts in E2U language would be convenient to better identify trends and test the representativeness of the results obtained, and such representativeness could be statistically tested to ensure the validity of the results (Corpas Pastor and Seghiri 2006). Likewise, some of the E2U principles discussed in the paper and included in present international guidelines have not been empirically tested as being easy for any language separately or may be only applicable to the English language—their applicability to other languages needs to be further explored in future studies. In this sense, a broader literature review encompassing literature from neighbouring fields, such as Psycholinguistics, could enrich the discussion of the results from a transdisciplinary perspective.

Despite these limitations, this corpus-based study opens new research opportunities. Let us underline three of them. The first one is that the study could be expanded to include other types of AD in Catalan, such as AD for museums, ballet, or opera AD. The second one is that the comparison of these results with corpus-based studies could be conducted for other languages by replicating the methodology adopted. The third and most important one is user validation. Our results suggest that Catalan AD is easy to understand, but this suggestion can only be confirmed by running experimental tests with users of E2U texts.

Acknowledgements

This research is part of the project Mediaverse, funded by the European Commission (H2020-EU2.1.1, ref. 957252). Blanca Arias-Badia is a member of TraDiLex, a research group recognised by the Catalan Government, under the SGR scheme (2021SGR00952). Anna Matamala is a member of TransMedia Catalonia, a research group funded by the Catalan Government, under the SGR scheme (2021SGR00077).

References

- **AENOR (Agencia Española de Normalización y Certificación)** (2005). *UNE 153020 Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías*. Madrid: AENOR.

- **Ardery, Gail** (1980). "On coordination in child language." *Journal of Child Language* 7, 305–320.

- **Arias-Badia, Blanca** (2021). "The audio description of sex in *The Affair*." Paper presented at the Advanced Research Seminar in Audio Description 2021, Barcelona, January 27. https://ddd.uab.cat/record/237607 (consulted 25.02.2023).

- **Arias-Badia, Blanca and 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, 251–270.

- **Arma, Saveria** (2011). *The language of filmic audio descriptions: a corpus-based analysis of adjectives*. PhD thesis. Università degli Studi di Napoli Federico II.

- **Baker, Paul** (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

- **Bassols, Margarida and Laura Santamaria** (2009). *L'audiodescripció en català*. Bellaterra: Publicacions de la Universitat Autònoma de Barcelona.

- **Bernabé-Caro, Rocío and Pilar Orero** (2019). "Easy to Read as multimode accessibility service." *Hermeneus* 21, 53–74.

- **Bernabé-Caro, Rocío and Pilar Orero** (2020). "Easier audio description. Exploring the potential of Easy-to-Read principles in simplifying AD." Sabine Braun and Kim Starr (eds) (2020). *Innovation in Audio Description Research*. New York: Routledge, 55–75.

- **Biber, Douglas** (1995). *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

- **Biber, Douglas *et al.*** (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

- **Corpas, Gloria and Miriam Seghiri** (2006). "El concepto de representatividad en la lingüística de corpus: aproximaciones teóricas y metodológicas." https://www.uma.es/hum892/publicaciones/corpas_seghiri_2006i.pdf (consulted 09.12.2022).

- **Cutts, Martin** (2020). *Oxford Guide to Plain English*. Oxford: OUP.

- **European Commission, Directorate-General for Translation** (2012). *How to write clearly*. Publications Office.

- **Fryer, Louise** (2016). *An introduction to audio description*. London: Routledge.

- **García Muñoz, Óscar** (2019). "*Diccionario Fácil*, una propuesta colaborativa para públicos con dificultades de comprensión lectora." Ignacio Blanco Alfonso, Luis Fernández-Martínez and Rebeca Suárez-Álvarez (eds) (2019). *Vulnerabilidad y cultura digital: riesgos y oportunidades de la sociedad hiperconectada*. Madrid: Dykinson, 327–345

- **Grodner, Daniel and Edward A.F. Gibson** (2005). "Consequences of the serial nature of linguistic input for sentential complexity." *Cognitive Science* 29(2), 261-290.

- **Halliday, Michael A.K.** (1985). *Spoken and Written Language*. Oxford: OUP.

- **Hermosa-Ramírez, Irene** (2021). "The hierarchisation of operatic signs through the lens of audio description." *Monografías de traducción e interpretación* 13, 184–219.

- **Hermosa-Ramírez, Irene** (2022). *La audiodescripción para ópera en España: studio desde la lingüística de corpus y la semiótica*. PhD thesis, Universitat Autònoma de Barcelona.

- **IFLA (International Federation of Library Associations and Institutions)** (2010). *Guidelines for easy-to-read materials*. IFLA Professional Reports, 120.

https://www.ifla.org/files/assets/hq/publications/professional-report/120.pdf (consulted 11.05.2022).

- **ILSMH-EA (International League of Societies for the Mentally Handicapped)** (1998). *Make it Simple. European guidelines for the production of easy-to-read information for people with learning disability for authors, editors, information providers, translators and other interested persons*. https://moodle2.units.it/pluginfile.php/303860/mod_resource/content/1/ILSMH%201998%20or%20Freyhoff%20Make%20it%20simple.pdf (consulted 11.05.2022).

- **Inclusion Europe** (2009). "Information for all: European standards for making information easy to read and understand." https://inclusion-europe.eu/wp-content/uploads/2015/03/2113-Information-for-all-16.pdf (consulted 11.05.2022).

- **ISO (International Organization for Standardization)** (2015). *ISO/IECTS 20071-21 Information technology – User interface component accessibility. Part 21: Guidance on audio description*. Geneva: International Organization for Standardization.

- **ISO (International Organization for Standardization)** (2023). *ISO/IEC 23859-1. Information technology – User interfaces – Part 1: Guidance on making written text easy to read and easy to understand*. Geneva: International Organization for Standardization.

- **Jiménez Hurtado, Catalina and Claudia Seibel** (2012). "Multisemiotic and multimodal corpus analysis in audiodescription: TRACCE." Aline Remael, Pilar Orero, and Mary Carroll (eds) (2012). *Audiovisual translation and media accessibility at the crossroads*. Amsterdam/New York: Rodopi, 409–425.

- **Lindholm, Camilla and Ulla Vanhatalo** (2021). "Introduction." Lindholm, Camilla and Ulla Vanhatalo (eds) (2021). *Handbook of Easy Language in Europe*. Berlin: Frank & Timme, 11–26.

- **López Sancio, Sergio** (2020). *Understanding dependencies in real time: A crosslinguistic investigation of antecedent complexity and dependency length*. PhD thesis, University of the Basque Country.

- **Lust, Barbara and Cynthia A. Mervis** (1980). "Development of coordination in the natural speech of young children." *Journal of Child Language* 7(2), 279–304.

- **Maaß, Christiane** (2020). *Easy Language-Plain Language-Easy Language Plus. Balancing comprehensibility and acceptability*. Berlin: Frank & Timme.

- **Maaß, Chistriane, Rink, Isabel and Silvia Hansen-Schirra** (2021). "Easy Language in Germany." Camilla Lindholm and Ulla Vanhatalo (eds) (2021). *Handbook of Easy Language in Europe*. Berlin: Frank & Timme, 191–218.

- **Maszerowska, Anna, Matamala, Anna and Pilar Orero** (eds) (2014). *Audio Description. New perspectives illustrated*. Amsterdam: John Benjamins.

- **Matamala, Anna** (2018). "One short film, different audio descriptions. Analysing the language of audio descriptions created by students and professionals." *Onomázein* 41, 185–207.

- **Matamala, Anna** (2019). "The VIW project. Multimodal corpus linguistics for audio description analysis." *Revista Española de Lingüística Aplicada* 32(2), 515–542.

- **Matamala, Anna** (2022). "Easy-to-understand language in audiovisual translation and accessibility: state of the art and future challenges." *X Linguae* 2022(2), 130–144.

- **Matamala, Anna and Pilar Orero** (2016). *Researching audio description. New approaches.* London: Palgrave.

- **Matamala, Anna and Pilar Orero** (2018). "EASIT: Easy Access for Social Inclusion training." Pierrette Bouillon, Silvia Rodríguez and Irene Strasly (eds) (2018). *Proceedings of the 2nd Swiss Conference on Barrier-free Communication* (BFC 2018). UNIGE Archive Ouverte, 68–70.

- **Matamala, Anna, Oncins, Estel·la and Pilar Orero** (2021). "EASIT." *Revista del Congrés Internacional de Docència Universitària i Innovació* 5. https://raco.cat/index.php/RevistaCIDUI/article/view/377503 (consulted 11.05.2022).

- **Mazur, Beth** (2000). "Revisiting Plain Language." https://www.plainlanguage.gov/resources/articles/revisiting-plain-language/. (consulted 11.05.2022).

- **Montolío, Estrella and Mario Tascón** (2020). *El derecho a entender: La comunicación clara, la mejor defensa para la ciudadanía*. Madrid: Prodigioso Volcán/Los Libros de la Catarata.

- **Navarrete, Marga** (2018). "The use of audio description in foreign language education." Laura Incalcaterra McLoughlin, Jennifer Lertola and Noa Talaván (eds) (2018). *Audiovisual Translation in Applied Linguistics: Educational Perspectives*. Amsterdam: John Benjamins, 129–150.

- **Oakes, Michael P.** (2012). "Describing a translational corpus." Michael P. Oakes and Meng Ji (eds) (2012). *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, 115–148.

- **Perego, Elisa** (2019). "Audio description: evolving recommendations for usable, effective and enjoyable practices." Luis Pérez-González (ed.) (2019). *The Routledge Handbook of Audiovisual Translation*. Milton Park: Routledge, 114–129.

- **Perego, Elisa** (2020). *Accessible communication: A cross-country journey.* Berlin: Frank & Timme.

- **PLAIN (Plain Language Action and Information Network)** (n.d.). "Plain language guidelines: Keep it conversational." https://www.plainlanguage.gov/guidelines/conversational/ (consulted 11.05.2022).

- **PLAIN (Plain Language Action and Information Network)** (2011). "Federal Plain Language Guidelines." https://www.plainlanguage.gov/media/FederalPLGuidelines.pdf (consulted 11.05.2022).

- **Plena Inclusión Madrid** (2023). "Diccionario Fácil." https://www.diccionariofacil.org/ (consulted 22.05.2023).

- **Puigdomènech, Laura, Matamala, Anna and Pilar Orero** (2007). "Bases per a un futur protocol d'audiodescripció per a l'àmbit català." Unpublished report. Barcelona: Universitat Autònoma de Barcelona.

- **ReadabilityFormulas** (n.d.). "The Gunning's Fog Index (or FOG) Readability Formula." https://readabilityformulas.com/gunning-fog-readability-formula.php (consulted 26.05.2023).

- **Remael, Aline, Reviers, Nina and Gert Vercauteren** (eds) (2015). *Pictures painted in words: ADLAB Audio Description guidelines*. Trieste: EUT.

- **Reviers, Nina, Remael, Aline and Walter Daelemans** (2015). "The language of Audio Description in Dutch: Results of a corpus study." Anna Jankowska and Agnieszka Szarkowska (eds) (2015). *New Points of View on Audiovisual Translation and Accessibility*. Bern: Peter Lang, 167–189.

- **Reviers, Nina** (2017). *Audio-description in Dutch: A corpus-based study into the linguistic features of a new, multimodal text type*. Unpublished PhD thesis. University of Antwerp.

- **Reviers, Nina** (2018). "Studying the language of Dutch audio description." *Translation and Translanguaging in Multilingual Contexts* 4(1), 178–202.

- **Salway, Andrew** (2007). "A corpus-based analysis of audio description." Jorge Díaz-Cintas, Pilar Orero and Aline Remael (eds) (2007). *Media for all. Subtitling for the deaf, audio description, and sign language*. Amsterdam: Rodopi, 151–174.

- **Snyder, Joel** (2014). *The visual made verbal: a comprehensive training manual and guide to the history and application of audio description*. Arlington, VA: American Council of the Blind.

- **Soler Gallego, Silvia** (2018). "Audio descriptive guides in art museums: A corpus-based semantic analysis." *Translation and Interpreting Studies* 13(2), 230–249.

- **Taylor, Christopher** (2015). "The language of AD." Aline Remael, Reviers, Nina and Gert Vercauteren (eds) (2015). *Pictures painted in words. ADLAB Audio description guidelines*. Trieste: EUT, 46–49.

- **Taylor, Christopher and Elisa Perego** (2021). "New approaches to accessibility and audio description in museum environments." Sabine Braun and Kim Starr (eds) (2021). *Innovation in Audio Description Research*. New York: Routledge, 33–54.

- **Vercauteren, Gert, Reviers, Nina and Kim Steyaert** (2021). "Evaluating the effectiveness of machine translation of audio description: the results of two pilot studies in the English-Dutch language pair." *Revista Tradumàtica* 19, 226–252.

- **Walczak, Agnieszka** (2016). "Foreign Language Class with Audio Description: A Case Study." Anna Matamala and Pilar Orero (eds) (2016) *Researching Audio Description*. London: Palgrave Macmillan, 187–204.

Data availability statement

The dataset is available at:
https://dataverse.csuc.cat/privateurl.xhtml?token=903a8ef1-c6ea-435c-a037-e3b6091387c5

Bios

Blanca Arias-Badia, PhD in Translation and Language Sciences (UPF, Barcelona), is a tenure-track lecturer at Universitat Pompeu Fabra. She is a member of the TraDiLex research group (UPF), an external collaborator of TransMedia Catalonia at Universitat Autònoma de Barcelona, and a

member of the network AccessCat. She is the principal investigador of the project UnivAc, funded by the Spanish Ministry of Science and Innovation, the Spanish Research Agency, and the European Union. Her publications include the single-authored monograph *Subtitling Television Series* (Peter Lang, 2020).

ORCID: https://orcid.org/0000-0003-1218-986X

Email: blanca.arias@upf.edu

Anna Matamala, PhD in Applied Linguistics (UPF, Barcelona), is a full professor at the Universitat Autònoma de Barcelona. She is the leader of Transmedia Catalonia research group and of the network AccessCat and has participated (DTV4ALL, ADLAB, HBB4ALL, ACT, ADLAB PRO, IMAC, TRACTION, Mediaverse) and led (AVT-LP, ALST, VIW, NEA, EASIT, RAD) funded projects on audiovisual translation and media accessibility. She received Joan Coromines Prize in 2005, APOSTA Award to Young Researchers in 2011, Dr. Margaret R. Pfanstiehl Memorial Achievement Award in Audio Description Research and Development in 2021.

ORCID: https://orcid.org/0000-0002-1607-9011

Email: Anna.Matamala@uab.cat

Appendix: Corpus aboutness results

The tables below show the absolute (A) and relative (R) frequencies of the 30 most frequent words in each AD script. For readability purposes, each table includes data of three of the films under study.

Deception	**A**	**R**	***Buried***	**A**	**R**	***Mamma Mia!***	**A**	**R**
ell	93	15.00	encenedor	25	9.21	donna	82	11.87
jonathan	88	14.19	mòbil	24	8.84	sophie	61	8.83
mirar	55	8.87	cap	23	8.47	fer	49	7.09
wyatt	37	5.97	mà	19	7.00	sam	47	6.80
anar	34	5.48	llum	19	7.00	rosie	47	6.80
noia	33	5.32	agafar	17	6.26	bill	38	5.50
vista	27	4.35	llanterna	16	5.90	tanya	37	5.35
carrer	27	4.35	fer	16	5.90	ell	37	5.35
cap	25	4.03	caixa	16	5.90	cap	37	5.35
fer	24	3.87	deixar	15	5.53	harry	36	5.21
sortir	23	3.71	sostre	14	5.16	tot	34	4.92
somriure	23	3.71	sorra	14	5.16	mirar	33	4.77
entrar	23	3.71	paul	13	4.79	mà	26	3.76
ros	22	3.55	apagar	13	4.79	anar	25	3.62
porta	22	3.55	peu	12	4.42	mentre	23	3.33
davant	22	3.55	amunt	12	4.42	sortir	20	2.89
observar	21	3.39	treure	11	4.05	sky	20	2.89
mà	20	3.22	ser	11	4.05	pati	19	2.75
agafar	19	3.06	mirar	11	4.05	somriure	17	2.46
treure	18	2.90	anar	11	4.05	escala	17	2.46
mirada	18	2.90	vara	10	3.68	home	16	2.32
llit	18	2.90	panxa	10	3.68	ballar	16	2.32
habitació	18	2.90	engegar	10	3.68	treure	15	2.17
creuar	17	2.74	cara	10	3.68	saltar	15	2.17
veure	16	2.58	tornar	9	3.32	moll	15	2.17
posar	16	2.58	terra	9	3.32	dinamos	15	2.17
passar	16	2.58	caure	9	3.32	asseure	15	2.17
vestíbul	15	2.42	petaca	8	2.95	alçar	15	2.17
sala	15	2.42	paret	8	2.95	posar	14	2.03
dona	15	2.42	enfocar	8	2.95	aigua	14	2.03

Midnight in Paris	**A**	**R**	***[REC3] Gnesis***	**A**	**R**	***The Contract***	**A**	**R**
gil	36	15.48	koldo	83	12.83	frank	76	15.99
ser	15	6.45	infectar	72	11.13	ray	74	15.57
ell	13	5.59	clara	66	10.20	home	46	9.68
cap	11	4.73	ell	54	8.35	cap	43	9.05
asseure	11	4.73	mirar	46	7.11	chris	33	6.94
cotxe	10	4.30	cap	35	5.41	bosc	23	4.84
anar	10	4.30	porta	32	4.95	mirar	21	4.42
mirar	9	3.87	anar	32	4.95	davis	21	4.42
home	9	3.87	fer	28	4.33	córrer	20	4.21
tot	8	3.44	haver	27	4.17	tot	19	4.00
fer	8	3.44	mà	24	3.71	sandra	19	4.00
aturar	8	3.44	adrián	22	3.40	fer	17	3.58
adriana	8	3.44	rafa	21	3.25	agafar	17	3.58
passejar	7	3.01	cuina	21	3.25	mà	16	3.37
costat	7	3.01	altre	21	3.25	avançar	16	3.37
carrer	7	3.01	saló	19	2.94	arma	15	3.16
acostar	7	3.01	ull	18	2.78	anar	15	3.16
treure	6	2.58	túnel	18	2.78	turner	14	2.95
porta	6	2.58	tiet	16	2.47	ser	14	2.95
on	6	2.58	passar	16	2.47	cotxe	14	2.95
inez	6	2.58	càmera	16	2.47	terra	12	2.53
haver	6	2.58	ser	15	2.32	noi	12	2.53
entrar	6	2.58	on	15	2.32	helicòpter	12	2.53
detectiu	6	2.58	dona	15	2.32	deixar	12	2.53
caminar	6	2.58	sang	14	2.16	avall	12	2.53
butxaca	6	2.58	ensangonar	14	2.16	aturar	12	2.53
vestir	5	2.15	apropar	14	2.16	treure	11	2.31
tornar	5	2.15	agafar	14	2.16	tornar	11	2.31
museu	5	2.15	veure	13	2.01	motxilla	11	2.31
mà	5	2.15	treure	13	2.01	johnson	11	2.31

Closer	**A**	**R**	***Law Abiding Citizen***	**A**	**R**	***Harry Potter and the Half-blood Prince***	**A**	**R**
ell	92	21.29	nick	87	15.51	harry	102	13.42
fer	37	8.56	clyde	69	12.30	noi	54	7.10
dan	35	8.10	cotxe	36	6.42	cap	48	6.31
alice	33	7.64	fer	32	5.70	mirar	45	5.92
larry	29	6.71	dunnigan	29	5.17	anar	44	5.79
anna	29	6.71	mirar	26	4.63	ron	40	5.26
mirar	22	5.09	ser	22	3.92	home	38	5.00
mà	22	5.09	haver	22	3.92	draco	36	4.74
cap	22	5.09	altre	21	3.74	tot	33	4.34
posar	21	4.86	tot	20	3.56	mà	32	4.21
aturar	19	4.40	mirada	20	3.56	fer	32	4.21
mirada	18	4.16	cap	20	3.56	dumbledore	28	3.68
mentre	18	4.16	cantrell	18	3.21	davant	28	3.68
acostar	18	4.16	davant	17	3.03	acostar	27	3.55
somriure	17	3.93	sortir	16	2.85	agafar	26	3.42
ull	16	3.70	presó	16	2.85	noia	23	3.03
davant	16	3.70	cella	16	2.85	hermíone	23	3.03
tornar	15	3.47	taula	15	2.67	gran	23	3.03
passar	15	3.47	sarah	15	2.67	altre	23	3.03
treure	14	3.24	sala	15	2.67	vareta	21	2.76
mig	14	3.24	porta	15	2.67	aigua	21	2.76
deixar	14	3.24	darby	15	2.67	ser	20	2.63
veure	12	2.78	creuar	15	2.67	obrir	20	2.63
noia	12	2.78	anar	15	2.67	girar	20	2.63
vista	11	2.55	agent	15	2.67	porta	19	2.50
taula	11	2.55	posar	14	2.50	on	19	2.50
porta	11	2.55	mà	14	2.50	tornar	18	2.37
casa	11	2.55	garza	13	2.32	sortir	18	2.37
damunt	10	2.31	ell	13	2.32	haver	18	2.37
asseure	10	2.31	passadís	12	2.14	ell	18	2.37