From image to text to speech : The effects of speech prosody on information sequencing in audio description

Hirvonen, Maija

2021-02

Hirvonen , M & Wiklund , M 2021 , ' From image to text to speech : The effects of speech prosody on information sequencing in audio description ' , Text & Talk , vol. 41 , no. 3 , pp. 309-334 . https://doi.org/10.1515/text-2019-0172

http://hdl.handle.net/10138/339505
https://doi.org/10.1515/text-2019-0172

unspecified
acceptedVersion

TEXT & TALK Submission

COVER SHEET

Author name(s): Maija Hirvonen (main author) & Mari Wiklund (second author)

Main author's affiliation (university and country): Tampere University and University of Helsinki, Finland

Main author's full institutional postal address: Kanslerinrinne 1, 30014 Tampere University, Finland.

Main author's contact email: maija.hirvonen@tuni.fi

Main author's ORCID ID: 0000-0002-7631-1310

Second author's affiliation (university and country): University of Helsinki, Finland

Second author's full institutional postal address: Unioninkatu 40, Department of Languages, P.O. Box 24, 00014 University of Helsinki, Finland.

Second author's contact email: mari.wiklund@helsinki.fi

Second author's ORCID ID: 0000-0002-1257-5341

Full article title: "From image to text to speech: The effects of speech prosody on information sequencing in audio description"

Short title (for running head): "From image to text to speech in audio description"

Word count (all inclusive): 8 365

Character count (all inclusive): 54 946

BIONOTES

Maija Hirvonen (PhD, 2014) is a tenure track professor (associate level) in German Language, Culture and Translation at Tampere University (Finland), and a visiting researcher at University of Helsinki (Finland). Her research focuses on accessibility (in particular audio description), multimodal and intermodal translation, multimodal interaction, and shared cognition.

Mari Wiklund (née Lehtinen) (PhD, 2009) works as a university lecturer in the field of French philology at the Department of Languages of the University of Helsinki (Finland). Her current research interests include e.g. prosody (French, Finnish), foreign accent, interaction of persons afflicted with autism, comprehension problems, conversational repairs, nonverbal communication, disfluencies of speech and asymmetric interaction. Her methodological background is mainly in conversation analysis and prosodic analysis.

From image to text to speech:

The effects of speech prosody on information sequencing in audio description

ABSTRACT

Given the extensive body of research in audio description – the verbal-vocal description of visual or audiovisual content for visually impaired audiences – it is striking how little attention has been paid thus far to the spoken dimension of audio description and its para-linguistic, prosodic aspects. This article complements the previous research into how audio description speech is received by the partially sighted audiences by analyzing how it is performed vocally. We study the audio description of pictorial art, and one aspect of prosody is examined in detail: pitch, and the segmentation of information in relation to it. We analyze this relation in a corpus of audio described pictorial art in Finnish by combining phonetic measurements of the pitch with discourse analysis of the information segmentation. Previous studies have already shown that a sentence-initial high pitch acts as a discourse-structuring device in interpreting. Our study shows that the same applies to audio description. In addition, our study suggests that there is a relationship between the scale in the rise of pitch and the scale of the topical transition. That is, when the topical transition is clear, the rise of pitch level between the beginnings of two consecutive spoken sentences is large. Analogically, when the topical transition is small, the change of the sentence-initial pitch level is also rather small.

Keywords: audio description, art, prosody, pitch, paragraph intonation, information structure, speech paragraph

1 Introduction

Audio description (AD) means translating visual and audiovisual content – such as art, cinema, television programs and theatre – into spoken verbal descriptions for the benefit of visually impaired audiences. AD requires vocal skills from the person delivering the description – a capacity for making meaning with the voice (Snyder 2008: 196). Although AD is a well-established subject of research in translation studies and linguistics, the vocal meaning making of *audio* description lacks investigation (Fryer 2016; Iglesias Fernández *et al.* 2014). Similarly, scholarly attention has centered around certain genres, the audio description of film being the most widely studied (see e.g. Fix 2005 and Maszerowska *et al.* 2014 for collections of textual analyses of film audio description, Mazur and Kruger 2012 for reception studies, and Fresno 2014 and Ramos 2015 for the study of cognitive and emotional effects in audio description). The AD of visual arts, which is the focus in this paper, is increasingly gaining more attention (e.g. De Coster and Mühleis 2007, Soler Gallego 2018a, Soler Gallego 2018b).

As opposed to audiovisual film and theater, visual arts lack the auditory dimension. They are "still images" (paintings, photographs, statues, etc.), meaning that the source text is static, while in cinema and theater it is dynamic, i.e. organized in temporal sequences and transforming in time. Previous research suggests that touch (Soler Gallego 2018b) and a multi-sensory experience (Neves 2012) can be used as auxiliary dimensions to complement the visual form. Nonetheless, the description of static source material places the describer in a position to decide the order of words and the 'path' of description. This article sets to study how information is structured via audio describer's voice. We report on a discourse-linguistic study and a prosodic

analysis of the vocal delivery of AD in works of art (mostly paintings).[1] This analysis complements the research that is largely concentrated on the linguistic, non-verbal aspects of AD and the AD of audiovisual content. The findings shed new insight into the practice of AD and the study of its reception by revealing ways in which the vocal delivery and the verbal formulation are contradictory and therefore potentially confusing to users.

The paper begins by reviewing the practice and research to audio describing visual art in Section 2.1. In the literature review (Sections 2.2 and 2.3), we pay special attention to the research on the vocal delivery in AD. This is followed in Section 3 by a description of our data and research methodology. Section 4 reports the data analysis, demonstrating and explaining the relation between pitch and information structuring with three examples. The paper ends with an extensive discussion (Section 5) and conclusions (Section 6).

2 Background

2.1 Some notes on the practice of audio describing visual art

Accessibility and services for special audiences have increased the demand for intermodal and intersemiotic translations, that is, conversions of visual and audiovisual materials to text and speech and vice versa. Television programs and theater shows now come with intralingual subtitles, sign-language interpretation as well as audio description to serve the Deaf and the hearing and visually impaired audiences. (For an overview, see e.g. Díaz Cintas *et al.* 2007.) Museums offer AD either as part of the guided tours, during which the description is performed live along with the guidance, or as pre-recorded descriptions in audio guides (e.g. Soler Gallego

[1] The authors wish to thank the Sara Hildén Art Museum and the Finnish National Gallery for allowing access to the data for this study.

2018b: 113). Today, museums are going online and also offer AD in their digital collections (e.g. Ateneum 2019 in Finland).

These different venues offer manifold possibilities to (audio) describe art and other objects. AD can be transmitted in speech either live or in a recording, or it can be a written text that is readable by text-to-speech software or a Braille text. AD speech can be spontaneous – performed as free speech – or it can be read out loud from paper. The person performing the AD may be a museum guide, an audio describer or a professional speaker. The Finnish National Gallery and its museum Ateneum (2019) present an interesting mixture: they have recorded a live audio-described guided tour and offer this recording as audio tracks on their website. Our data are examples of recorded audio descriptions that Finnish museums offer or have offered on their websites, and they are performed by a female audio describer/museum guide and by a male professional speaker.

2.2 Previous research on the audio description of art

While previous research on the AD of art (De Coster and Mühleis 2007; Neves 2012; Soler Gallego 2018a and 2018b) has discussed the intersemiotic rendering from visual to verbal representation, the second stage of transformation – from text to speech – has been largely overlooked. Soler Gallego (2018b) conducted a corpus analysis of audio described art and defined three levels of semantic specificity:

> "the translation of the work as a whole (Level 1), the translation of the content and formal components it is made of (Level 2), and the translation of the specific properties of the elements described in the previous level (Level 3)" (Soler Gallego 2018b: 117).

In addition to the verbal text, the use of other communication modes has been discussed, such as transforming the visuals into a tactile representation (Soler Gallego 2018b: 120–121) or to a multisensory representation that employs touch, hearing and material objects (Neves 2012). These modes come with distinct properties from language: a tactile image for instance allows the blind user to perceive the composition of a work of art holistically, with two hands simultaneously, as opposed to the linear organization of language (see Hirvonen 2014: 24). Voice, including tone, rhythm and speech modulation, can work together with music and sound effects to produce emotions in listeners (Neves 2012: 290). As we will show in our analysis, voice – and speech prosody in particular – provides an effective device for creating and maintaining coherence in AD (see also Soler Gallego 2018b: 120).

2.3 Previous research on the vocal delivery in AD

Speech is central to AD because, in the absence of visual perception, verbal descriptions (words, sentences, etc.) are delivered in an audible form in order to be received by the blind user. Previous studies have investigated the reception of AD delivery (user preferences, perceptions, etc.) (e.g. Szarkowska and Jankowska 2012; Iglesias-Fernández *et al.* 2014). In this article, we deal with delivery *per se*, with its linguistic and acoustic features.

Apart from spontaneous descriptions which can occur in any situation where blind and sighted people interact, AD is usually a hybrid form of writing and speaking – a specific text type constituting of a written text that is read out loud (TROL). TROL has features of both written-ness and spoken-ness (Gutenberg 2000). The written-ness causes the spoken utterances to become longer and more complex than they typically are in spontaneous speech (Gutenberg 2000: 579–580). As regards German, it has been noted that emphases (e.g. stress on individual

words) occur more often in TROL than in spontaneous speech. It is also characteristic of German TROL that there is an abundance of nominalization and the use of attributes (Gutenberg 2000: 579–580). Poethe (2005) found that the German AD of film is just that – full of nominalization and attributes. This trend to pack information in AD densely has to do with meeting the objective of information abundancy (Kluckhohn 2005). Yet the double functionality of AD – it is supposed to be easy to follow yet rich in information – may produce a contradictory effect and render the speech unintelligible.

Guidelines directed at the practice of AD instruct describers to use a neutral voice and keep the information in each sentence limited to one issue and the speech pace at 160 words/minute (Remael *et al.* 2015: 48). Voicing is regarded as an important part of the production process (e.g. the ADLAB (Audio Description: Lifelong Access for the Blind) guidelines by Remael *et al* 2015) and some instructions are given. The relevance of voice quality is brought up but in a rather general sense, such as what kind of voice fits with what film genre and style. Means of prosodic realization are mentioned briefly and indirectly. Audio describers should indicate in the script when the speaker needs to speed up their oral delivery to accommodate it in the soundtrack, and they may give advice for using intonation: "more empathetically for an emotional fiction film, more newscaster style for a documentary, or adapted for text on screen" (Remael *et al.* 2015: 57).

A recent practical guide for AD (Fryer 2016) goes into more detail about the use of voice. Fryer categorizes the delivery of AD along with writing and script preparation as the three main areas of "audio description skills". The delivery relates to "the supra-linguistic aspects of speech [that] convey meaning through stress, pitch, tempo, dynamic range and, especially, the way the words are segmented […]" (Fryer 2016: 87). For instance, the rhythmic structure of language

realized through phonetic stress and segmentation are important because they direct the listeners' attention and help them to interpret meaning units from speech (Fryer 2016: 90). In the present study, we set out to find out how exactly the segmentation occurs in AD.

3 Data and methodology

3.1 The AD data

In this study, we analyze audio descriptions of art works from two Finnish museums: The Ateneum Art Museum in Helsinki, and the Sara Hildén Art Museum in Tampere. Most of the works are paintings (oil or gouache on canvas), but a few are drawings (Miró is charcoal/pastel and Picasso is charcoal). The audio descriptions are in Finnish and they currently are (Ateneum[2]) or have been (Hildén) accessible on the websites of the museums. The audio descriptions are voiced by two different persons (a female audio describer at Ateneum and a male professional speaker at Hildén). The Ateneum collection entails six pieces of classical Finnish art. The collection from Hildén, in contrast, describes seven works of art and exhibits a selection of international classics. A detailed account of the data can be found in Table 1, which also shows the differences in the length of AD in the collections.

Table 1: List of data

[2] Listen to the Ateneum's AD here: https://ateneum.fi/opastukset/kuvailutulkkaukset/

	Work of art (Ateneum)	Length of AD (min:sec)	Work of art (Sara Hildén)	Length of AD (min:sec)
1	Ferdinand von Wright: *Taistelevat metsot* ('The Fighting Capercaillies') (1886)	06:12	Pierre Bonnard: *Seisova alaston sinisen kylpyaltaan edessä* ('In the Bathroom') (1907)	02:55
2	Otto Mäkilä: *Kesäyö* ('Summer Night') (1938)	06:07	Giorgio de Chirico: *Trubaduuri* ('Troubadour') (1940)	03:02
3	Tyko Sallinen: *Pyykkärit* ('The Washerwomen') (1911)	09:59	Paul Delvaux: *Kesä* ('Summer') (1938)	02:57
4	Helene Schjerfbeck: *Toipilas* ('The Convalescent') (1888)	08:54	Paul Klee: *Tapaus satamassa* ('Harbour scene') (1923)	02:59
5	Eero Järnefelt: Raatajat rahanalaiset (a.k.a. Kaski) ('Under the Yoke' a.k.a. 'Burning the Brushwood') (1893)	09:30	Fernand Léger: *Monivärinen kukka* 'Colourful flower' (1937)	03:07
6	Albert Edelfelt: *Pariisin Luxembourgin puistossa* ('Luxembourg Gardens, Paris') (1887)	11:52	Joan Miró: *Olentoja yössä* ('Creatures in the night') (1942)	02:50
7			Pablo Picasso: *Lasi ja viulu* ('Glass and violin') (1912-13)	03:20

As can be seen in Table 1, the audio descriptions vary in length quite considerably. The descriptions of Ateneum are 2–3 times longer than those of Hildén. The length of the Ateneum descriptions is partly due to the fact that they include a general "guidance" to the work as well: for the last 1–2 minutes, the describer gives background information concerning the artwork or the artist. This, however, does not explain the whole difference; some of the differences are related to the content or the prosodic realization of the descriptions. We measured the speech rates of the two describers: the female describer's speech rate is 3.6 syllables per second, and

the male describer's speech rate is 3.9 syllables per second. Thus, both describers speak rather slowly,[3] and there is no notable difference in the speech rates of the two describers.

3.2 The methodology

3.2.1 The concept of speech paragraph

It is often said that the way we speak conveys as many meanings as the words we use (Couper-Kuhlen 2000: 2). A large part of this "way we speak" consists of prosodic features. In phonetics, *prosody* encompasses phenomena related to pitch (voice fundamental frequency, f_0), duration (timing), loudness (intensity), speech rate, speech rhythm and phrasing (Crystal 1969, 1980). According to Couper-Kuhlen (2000: 2), such para-linguistic features of speech as breathings, creaky voice, nasal voice and whispering cooperate so closely with prosody that they can also be included in prosodic features when the term is used in the broad sense. In this study, we use the term 'prosody' in this broad sense, even if our analyses mainly concern the role of pitch changes.

An important concept in our study is a 'speech paragraph' (or a 'spoken paragraph', Wichmann 2000). A speech paragraph starts on a remarkably high pitch level. This phenomenon has also been called 'topic reset', even if it does not always necessarily indicate a clear change of topic (Wichmann 2000: 25). We will refer to this phenomenon in this study simply by the term 'sentence-initial high pitch' (SIHP). The high pitch occurs on the onset – that is, on the *first accented syllable* of a spoken sentence. Even if the phenomenon is called 'sentence-initial high pitch', the syllable carrying the high pitch is not always necessarily the first syllable of the

[3] According to Koskela (2013), the mean speech rate of Finnish-speaking adults is 4.96 syllables per second.

spoken sentence. Nafá Waasaf (2007) has shown that in simultaneous interpreting both the speakers and the interpreters treat a SIHP as a sign that indicates the beginning of a new topical unit. The same applies to speech-to-text interpreting (Wiklund 2014). In this study, we are interested in finding out if the occurrences of the SIHP phenomenon are related to topical transitions also in AD, i.e. how a written-textual phenomenon reflects in the spoken text (cf. Fryer 2016: 87).

Previous studies (e.g. Sluijter and Terken 1993; Wichmann 2000) show that spoken sentences are typically linked internally via an overriding declination line. That is, there is a 'supradeclination' between the beginnings of spoken sentences occurring inside the same speech paragraph: a new spoken sentence generally starts on a lower pitch level than the preceding one. This applies especially to data coming from highly controlled experimental settings (e.g. Sluijter and Terken 1993). In data coming from naturally occurring situations, the supradeclination is not systematic (Wichmann 2000: 121). Indeed, according to Wichmann (2000: 121), the supradeclination constitutes an "envelope" within which linguistically motivated variation operates. For example, the information structure of a speech paragraph plays a role in determining the level at which each spoken sentence starts. It is possible, for example, that reformulations and precisions start on a rather low level, and after this, a spoken sentence that continues the actual topic starts again on a slightly higher level (Wichmann 2000).

The pitch on the first accented syllable of a speech paragraph is extra high. Unaccented syllables before the first accent also tend to be high (Couper-Kuhlen 2006; Wichmann 2000). The end of a speech paragraph is generally indicated by an extra-low pitch, close to the

speaker's baseline, and often by a noticeable pause (Couper-Kuhlen 1986, 2006; Wiklund 2014).[4]

We use the term 'spoken sentence' to refer to a unit which, despite its name, is not in the first instance a syntactic unit but rather a prosodic one (Couper-Kuhlen 2006; Wichmann 2000). The beginning of a spoken sentence is indicated by a high pitch on the first accented syllable of an intonation phrase.[5] That is, the purpose of this paper is to show where in the descriptive text *a reader chooses to indicate a shift*, whether in line with syntax or not. The end of a spoken sentence, in turn, is signaled by a falling pitch starting on or from the last accented syllable of an intonation phrase and reaching a low point in the speaker's voice range (Couper-Kuhlen 2006; Wichmann 2000). Canonically, the accents in the spoken sentence form a pitch line that gradually descends, or declines, throughout the unit (Couper-Kuhlen 1986, 2006). Laver (1994) has postulated a similar line of declination for amplitude. Spoken sentences may consist of several intonation phrases (or tone groups): if there are several, the groups are linked by a single declination line for pitch and amplitude (Couper-Kuhlen 2006; Wichmann 2000). Spoken sentences do not always correspond to a syntactic/orthographic sentence; the same prosodic pattern is also used, for example, in titles and other noun phrases. In our data, a spoken sentence most often corresponds more or less to a syntactic/orthographic sentence because the

[4] In Wiklund's (2014) study, however, speech paragraphs also rather often ended in a pitch rise when the speaker was female. Wiklund's data come from Finnish speech-to-text interpreting situations.

[5] As our data are in Finnish, an 'intonation phrase' corresponds here to a 'minor intonation unit' as defined by Aho (2010: 39-42). The boundaries of a minor intonation unit can be marked with changes in pitch, volume, speech rate or quality of voice, or with other phonetic features occurring alone or together. The duration of a minor intonation unit is usually 1-2 seconds. A minor intonation unit does not necessarily always have semantic content, but it typically has a pragmatic value.

data consist of written texts that are being read aloud. Sometimes a spoken sentence may, however, consist of a single noun phrase. This is the case, for example, in the titles of pieces of art being described.[6]

In Aho's (2010: 35) classification concerning spontaneous Finnish speech, a spoken sentence corresponds to a 'major intonation unit' (*laaja intonaatiojakso*). Its duration is generally 5–10 seconds. A major intonation unit typically includes things that are closely related to each other, and it almost always ends in a pause. The pauses between major intonation units are often rather long, and they clearly indicate the boundaries between units. Generally, a major intonation unit starts on a high pitch level, and the pitch curve gradually declines towards the end of the unit. In the beginning of a major intonation unit, the amplitude usually becomes remarkably larger, and then it becomes smaller towards the end (Aho 2010: 36). Sometimes the mere shape of the amplitude may be enough to signal the boundaries between intonation units. Chafe's notions of 'sentence' (Chafe 1994: 140) or 'center of interest' (Chafe 1980: 26) are also reminiscent of Wichmann's (2000) spoken sentence.

3.2.2 The analytical procedure

Our study is based on Wichmann's (2000) approach, a combination of discourse analysis and intonation studies. In Wichmann's words (2000: 2) words, this approach makes use of "both auditory and instrumental analysis, thus taking into account what the listener hears and what the computer can measure." Both auditorily transcribed texts and the corresponding sound recordings were used. In the transcriptions, we applied the standard conversation analytic transcription conventions (see e.g. Arminen 2016 and Appendix). The verbal data of the AD

[6] The perception of sentence and paragraph boundaries is approached in Kreiman (1982).

was analyzed in terms of information structure (Lambrecht 1994). Particular attention was paid to transitional sequences, that is, the verbalizations at the end and beginning of speech paragraphs.

We used the speech analysis program Praat for the acoustic analyses of the data (Boersma & Weenink 2017). The pitch level (f_0) was measured at the moment of the production of the onset (first accented syllable) of each spoken sentence, and it was measured in semitones (st) with regard to 100 hertz (re 100 Hz). We chose to use semitones because they clearly show the change between the pitch levels of different units. An example of a pitch curve generated by Praat during the production of a spoken sentence ('metsäaukealla leijuu vielä aamu-usva', *in a clearing, the morning mist still floats*) is given in Figure 1.[7]

[7] The pitch curve has been corrected manually.

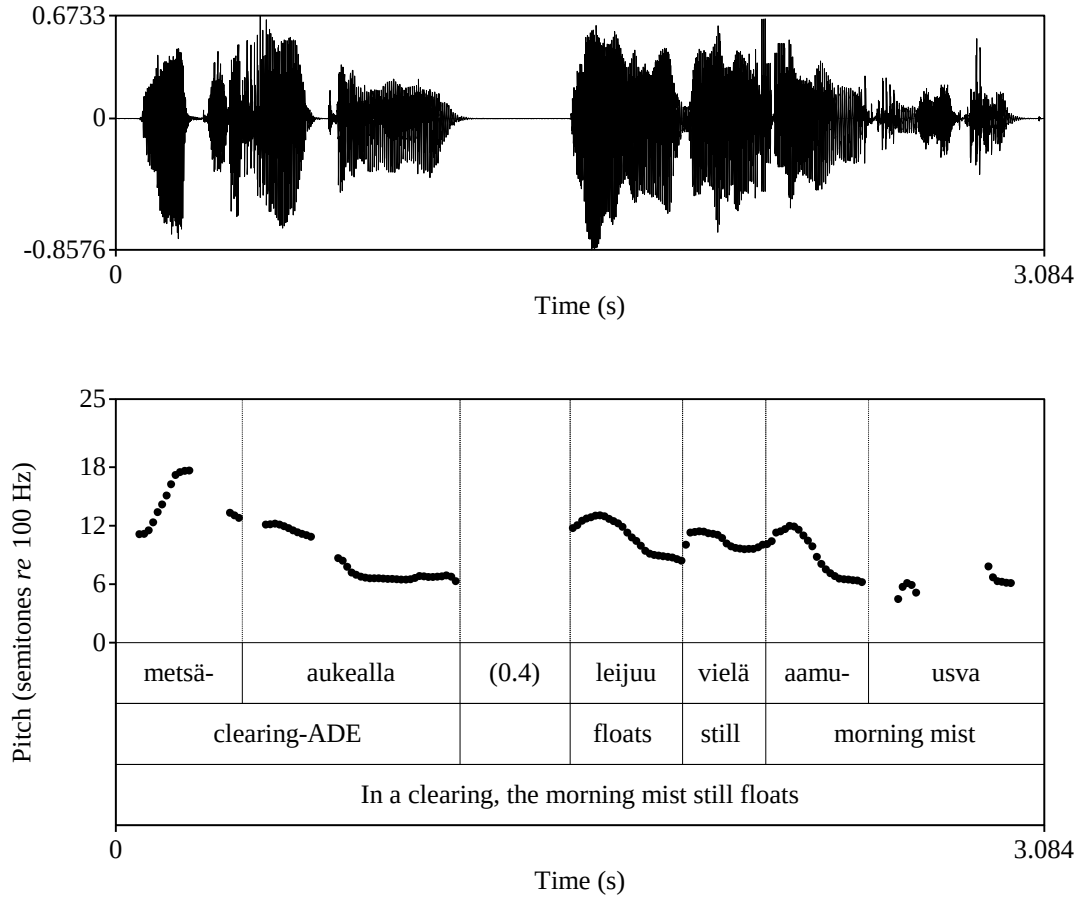

Figure 1: Pitch curve during the production of a spoken sentence

4 Data analysis: Speech paragraphs in audio description

In our data analysis, we were interested in finding out how the prosody of the audio describer's voice affects the delivery of information in AD, i.e., how is the AD structured in speech paragraphs and what is the relation of these paragraphs to the verbal information structuring? Are they mutually supportive, which would probably be advantageous to the listener, or contradictory, which is likely to lead to misunderstanding and lack of ease in listening?[8]

[8] The scope of the study is limited in that it applies a well-known feature of prosody (topic initial pitch reset) to a rather small number of audio descriptions of paintings. Not all paintings lend themselves to an obvious

In the first example, a female voice presents the AD of Ferdinand von Wright's masterpiece, *The Fighting Capercaillies* (1886). The numbers on top of the lines indicate the pitch level during the production of the first accented syllable of each spoken sentence. Stressed syllables are written in capital letters, and an arrow pointing upwards (↑) indicates a raised pitch level.

The first two examples (1 and 2) illustrate cases in which the prosodic (speech) and the thematic (words) realization of the description are symmetrical in the sense that the prosodic structuring – in the form of speech paragraph distribution – corresponds to a coherent thematic structuring. The last example (3) represents the opposite case, showing how the prosodic realization contrasts with the thematic progression. The examples are taken from the very beginning of the AD.

Example (1)

Ferdinand von Wright: *The Fighting Capercaillies* (1886) (Ateneum 1 / 00:24–01:02)

1ST SPEECH PARAGRAPH: GENERAL VIEW OF THE CONTENTS AND ATMOSPHERE

narrative structure. Thus, this analysis is selective and only indicative of how prosody is relevant to such readings.

14.9 st (re 100 Hz)

01 ↑METsäaukealla (0.4) leijuu vielä aamu-usva. (1.4)

clearing-ADE floats still morning mist

In a clearing, the morning mist still floats.

11.6 st

02 ↑SAMmaleen peittämälle AUkealle

moss-GEN covered clearing-ALL

on pysähtynyt vastakkain kaksi metsoa. (0.7)

have stopped face to face two capercaillies

In the clearing covered with moss,

two capercaillies have stopped face to face.

10.2 st

03 ne ↑TUIjottavat tiukasti toisiaan. (1.0)

they stare intensely at each other

They stare intensely at each other.

After telling the name and the size of the painting (not transcribed here), the first spoken sentence (*In a clearing, the morning mist still floats*, line 01) starts on a remarkably high pitch level. This spoken sentence starts a speech paragraph in which the speaker gives an overview of the contents and the atmosphere of the piece of art. The next spoken sentence (*In the clearing covered with moss, two capercaillies have stopped face to face*, line 02) starts on a 3.3 semitones lower pitch level than the preceding spoken sentence. In this spoken sentence, the speaker mentions the two main figures in the painting for the first time. The last spoken sentence (*They stare intensely each other*, line 03) describes the relationship between the main figures. It starts 1.4 semitones lower than the preceding spoken sentence. The thematic progression is as follows: *in a clearing* > *in the clearing … two capercaillies* > *they*. The first spoken sentence begins by introducing new information, a location, which becomes the known information in the second spoken sentence. The action that is described to take place in that location is being continued in the third spoken sentence as the pronoun *they* refers back to the

two capercaillies. Thus, the topic changes from the location to the agents performing action in that location.

2ND SPEECH PARAGRAPH: TOPICAL TRANSITION TO THE BACKGROUND

14.7 st

04 vain ↑MUUtaman askeleen PÄÄhän¿ (.)
only few steps away
metsikön REUnaan, (0.4)
wood-GEN border
on kuvattu NAArasmetso. (1.5)
is depicted hen capercaillie
Only a few steps away, at the edge of the forest,
is a capercaillie hen.

13.8 st

05 ↑METsojen Ympärillä¿ (0.3)
capercaillies-GEN around
LEvittäytyy SUUri mäntymetsä, (0.2)
spreads out large pine forest
MUStikanvarpuineen ja (.) KAnervikkoineen. (1.0)
twigs of blueberries-COM and heathlands-COM
Around the capercaillies, a large pine forest spreads out
with twigs of blueberries and heathlands.

11.7 st

06 ↑TAUStalle avautuu LAAksomainen (.)
background-ALL spreads out valley-like
jylhä maisema. (0.6)
wild scenery
In the background,
a valley-like wilderness scenery spreads out.

10.1 st

07 ↑JÄRven takana (0.2) siintää vuoria. (1.6)
lake-GEN behind looms mountains
Beyond a/the lake, mountains loom.

Then there is a change of speech paragraph (line 04). The pitch level rises 4.5 semitones with regard to the beginning of the previous spoken sentence (line 03). A rise of this order indicates

clearly the beginning of a new entity. Indeed, the describer proceeds to describing a new entity: a new location with a new agent (*Only a few steps away, at the edge of the forest, is a hen capercaillie*). After this, the pitch level falls only 0.9 semitones in the beginning of this spoken sentence compared with the beginning of the preceding spoken sentence, so that both of these sentences can be interpreted as new speech paragraphs. A new start is relevant because more entities are being described: the pine forest around the capercaillies (line 05), and the mountain scenery with lakes in the background (lines 06–07). However, these spoken sentences (lines 04 and 05) can be also treated as belonging to the same speech paragraph.

Thematically, this speech paragraph is rather complex. The description of the hen capercaillie has a similar syntactic realization as in line 02, so that the listener might expect the description to give more details about this entity. Instead, the describer introduces new information, albeit in a known context (*around the capercaillies, a large pine forest…* line 05, and *in the background* [of the capercaillies and the forest], *a valley-like wild scenery*… line 06). Indeed, the first syllables of the words *around* (‘ympärillä’) and *large* (‘suuri’) (line 05) are produced on the same pitch level as the onset (‘met-‘), which reflects the complex information structure of the spoken sentence. *Behind a lake* (line 07) seems a new element but considering that the given context is a ‘valley’ location, a ‘lake’ element is mentally accessible information and thus not totally new (Lambrecht 1994) as lakes are typical parts of valleys. AD makes abundant use of such discourse and referential devices in order to compress the description and make the information run smoothly (Hirvonen 2012).

The further in the background the description goes, the lower the sentence-initial pitch level is: the capercaillie hen a few steps away, the forest around the birds, and finally the scenery in the

background. In other words, the declination of pitch here reflects quite accurately the transition of directing attention from the main figures to other objects and areas.

3RD SPEECH PARAGRAPH: TOPICAL TRANSITION TO STYLE

16.6 st

08 ↑TAIStelevat METsot on VAlokuvantarkka ja,
fighting capercaillies is photographic and
(0.3) Yksityiskohtaisesti toteutettu. (1.0)
in detailed manner accomplished
The Fighting Capercaillies is photographic
and accomplished with great detail.

In the beginning of the third speech paragraph (line 08), there is a remarkable pitch rise: the pitch level rises as much as 6.5 semitones with regard to the beginning of the previous spoken sentence. Indeed, a strong prosodic cue is needed here because the topical transition is as significant: in this speech paragraph the description changes from content to the style of the painting.

Figure 2: *The Fighting Capercaillies*, Ferdinand von Wright, 1886 (photo by Finnish National Gallery, reproduced here under the CC license[9])

In the second example, a male voice presents Fernand Léger's piece of art, *Colourful Flower* (1937).

Example (2)

Fernand Léger: *Colourful Flower* (1937) (Sara Hildén nr 5; 00:26–01:13)[10]

1ST SPEECH PARAGRAPH: GENERAL VIEW OF THE CONTENTS

[9] https://www.kansallisgalleria.fi/fi/object/389906 (accessed 21 July 2020)
[10] The image of the artwork can unfortunately not be reproduced here due to copyright reasons. The image can, however, be found on the internet: https://www.flickr.com/photos/tvbrt/4835321748 (accessed 21 July 2020).

12.7 st (re 100 Hz)

01 ↑MAAlaus KOOStuu ERImuotoisista SUUrista (.)
painting consists of differently-shaped big
ja TOIsiinsa lomittuvista Osista. (1.3)
and to each other overlapping parts
The painting consists of big, overlapping parts
of different shapes.

3.0 st

02 ↑Osissa on käytetty PÄÄasiassa MUStaa ja valkoista, (0.4)
parts-INE has been used mostly black and white
sekä RUNsaasti PUnaista (.) sinistä (.)
as well as plenty of red blue
keltaista (.) ja vihreää. (1.5)
yellow and green
In the parts, mostly black and white as well as plenty of
red, blue, yellow and green have been used.

Even before this extract, the describer has told the name and the size of the piece of art. In the first speech paragraph, the describer starts to give a general view of the contents. The first spoken sentence (*The painting consists of big, overlapping parts of different shapes*, line 01) starts on a high pitch level. After this (line 02), he gives more information on the *parts* that were introduced in the preceding spoken sentence, so the thematic progression is clear. The pitch level falls as much as 9.7 semitones in the beginning of this second spoken sentence. The large fall of the sentence-initial pitch level illustrates the relationship between the spoken sentences – the fact that the second one elaborates on the first one.

2ND SPEECH PARAGRAPH: THE GENERAL IMPRESSION

14.7 st

03 ↑KOkonaisvaikutelma on VÄrien osalta

general impression is colours concerning

RÄISkyvä ja VOImakas. (2.2)

effervescent and strong

The general impression concerning the colours is effervescent and strong.

13.1 st

04 ↑TEOKsen TAUSta on VAAlea VIhertävän harmaa, (.)

artwork-GEN background is light greenish grey

hieman LIkaisen värinen. (1.7)

a bit dirty-GEN colored

The background of the artwork is light, greenish grey, a bit smudged.

8.4 st

05 ↑HIUkan teoksen keskiosan vasemmalla puolella; (.)

a little bit artwork-GEN center left-ADE side-ADE

on SUUrikokoinen KUkankaltainen kuvio. (2.2)

is large-sized flower-like figure

A little bit left from the center of the artwork

is a large-sized flower-like figure.

In the beginning of the second speech paragraph (*The general impression concerning the colors is effervescent and strong*, line 03), the pitch level rises remarkably. It rises as much as 11.7 semitones compared with the beginning of the preceding spoken sentence. This rise indicates a change: here, it is a change in point of view. The speaker has enumerated different colors that have been used and describes the overall impression created by the colors. Thematically, however, this spoken sentence could still be part of the preceding speech paragraph as it elaborates on the topic of colors. Given this contradiction, the rise of pitch may indicate here rather an emphasis on the word *KOkonaisvaikutelma* 'general impression' than a beginning of a new paragraph. This interpretation is supported by the prosodic and thematic presentation of the following spoken sentence (line 04) where the speaker moves on to describing the colors of the background. The pitch level falls only 1.6 semitones in the beginning of this spoken

sentence compared with the beginning of the preceding one, so that both of these sentences can be interpreted as new speech paragraphs.

In the third spoken sentence of the speech paragraph (line 05), the speaker moves on to a flower-like figure. In the beginning of this spoken sentence, the pitch falls 4.7 semitones. Overall, the focus of attention becomes more and more precise as the speech paragraph proceeds: he starts from the overall impression, moves on to the description of the background, and finally focuses on one aspect presented in the artwork.

3RD SPEECH PARAGRAPH: TOPICAL TRANSITION TO THE CENTRAL FIGURE

10.8 st

06 ↑KUviossa voi nähdä KUkan sijasta
figure-INE can see flower-GEN instead
esimerkiksi MEritähden tai POTkurin muodon. (1.1)
for example starfish-ACC or propeller-ACC shape-ACC
In the figure, instead of seeing a flower,
one might see, for example,
the shape of a starfish or propeller.

8.4 st

07 ↑TEos VAIkuttaisi saaneen nimensä (0.3)
artwork seems to have gotten its name
MOnivärinen KUKka (.) TÄStä MUOdosta. (2.4)
multicolored flower this-ELA shape-ELA
The artwork seems to have gotten its name Colourful Flower
from this shape.

In the third speech paragraph, the speaker elaborates on the flower-like figure which he introduced in the last spoken sentence of the preceding speech paragraph. In the first spoken sentence (line 06), he tells about the shape of the figure and then (line 07) gives his interpretation of the meaning of this figure. The pitch level rises only 2.4 semitones in the

beginning of this speech paragraph, which creates an impression of a rather small topical transition. Indeed, the transition is small because the figure that is being described has already been introduced at the end of the preceding speech paragraph. The point of view does not actually change here but it gets more precise: the central figure in the piece of art is now being discussed in more detail.

In the third example, the male voice presents Giorgio de Chirico's painting *Troubadour* (1940).

Example (3)

Giorgio de Chirico: *Troubadour* (1940) (Sara Hildén nr 2; 00:15–01:00)[11]

1ST SPEECH PARAGRAPH: GLOBAL VIEW OF THE PIECE OF ART

[11] The image of the artwork cannot be found on the internet. Yet images of similar artworks by the artist are findable, such as this one: https://www.flickr.com/photos/mbell1975/43610378920 (accessed 21 July 2020).

11.9 st (re 100 Hz)

01 ↑TEOSta REUnustaa LEveät PUUkehykset; (0.4)
artwork-PAR edges wide wooden frames
joiden KESKIosa on SUKlaanRUSkea ja REUnat
whose middle part is chocolate brown and sides
KIILtävän kultaiset. (2.1)
shiny-GEN golden
The artwork is edged with wide wooden frames
whose middle part is chocolate brown
and the sides shiny golden.

5.9 st

02 ↑MAAlausta hallitsee sen KESkelle sijoittuva
painting-PAR dominates its middle locate-1st-PART-NOM
SUUri pelkistetty IHMIShahmoinen nukke; (0.7)
big simplified human-figured doll
TRUbaduuri. (1.4)
troubadour
The painting is dominated by a big, simplified, human-like
doll located in the middle, a/the troubadour.

Once again, the description has begun by stating the name and the size of the piece of art. Then, in the first paragraph of this example, the describer gives a global view on the piece of art. The first spoken sentence (*The piece of art is edged with wide wooden frames whose middle part is chocolate brown and the sides shiny golden*, line 01) starts on a high pitch level. The second spoken sentence (line 02) starts 6.0 semitones lower than the first one. Thus, they clearly belong to the same speech paragraph. In the second spoken sentence, the describer introduces the central figure, the troubadour. Thus, at this point, the AD moves from the appearance of the *artwork* to describing the contents of the *painting*. Indeed, the AD of art typically addresses different levels of visual art (Soler Gallego 2018b: 117). The fact that the pitch level decreases remarkably in the beginning of the second spoken sentence compared with the first one illustrates this large topical transition. However, as both spoken sentences are thematically

related to the description of the global view of the piece of art, they are related to each other. Therefore, it is logical that they belong to the same speech paragraph.

2ND SPEECH PARAGRAPH: THE CENTRAL FIGURE

13.7 st

03 se ↑MUIStuttaa TAIteilijoiden
it resembles artists'
MALlinaan käyttämää, (.) PUISta JÄsennukkea. (1.3)
model-ESS use-3rd-PART-PAR wooden doll with limbs
It resembles a wooden doll with limbs
that artists use as a model.

8.5 st

04 ↑NUKke seisoo LEveässä HAAra-Asennossa (.)
doll stands broad-INE legs-apart position
lähes KOHtisuorassa, (0.3) TEoksen KATsojaa kohden. (1.0)
almost perpendicular artwork-GEN viewer towards
The doll stands with its legs wide apart,
almost facing the viewer of the artwork.

7.9 st

05 sillä on ↑LEVEät hartiat; (1.0)
it-ADE has broad shoulders
It is broad-shouldered.

5.0 st

06 ↑KÄsivarret (0.3) ja KÄMmenet
arms and palms
PUUTtuvat kokonaan. (1.1)
lack completely
Arms and palms are missing completely.

The second speech paragraph of the extract starts with the spoken sentence *It resembles a wooden doll with limbs that artists use as a model* (line 03). The onset of this spoken sentence is produced as much as 7.8 semitones higher than the onset of the preceding spoken sentence. Thus, it is clear that this unit starts a new speech paragraph. The topical transition, however,

presents a contradiction. Even though this speech paragraph overall is a coherent entity – a general description of the central figure in the painting – it begins with an anaphoric reference (the pronoun *it*) to the doll figure introduced in the preceding speech paragraph. Generally, new referents, which are not known in the context of the discourse, are introduced with sentence-initial high pitch in new paragraphs, and the known referents initiate spoken sentences in the middle of speech paragraphs (see Example 1). The second spoken sentence that elaborates the description of the doll (*The doll stands with its legs wide apart, almost facing the viewer of the artwork*, line 04) starts 5.2 semitones lower than the first spoken sentence. The third spoken sentence, in turn, starts on a level 0.6 semitones lower than the preceding one. In this unit, the speaker states that the central figure (the doll) is broad-shouldered (line 05). In the last spoken sentence of the paragraph, the speaker states that the arms and the palms of the doll are missing completely (line 06). This spoken sentence starts 2.9 semitones lower than the preceding one. Thematically, thus, the point of view gets more and more precise: first the description is focused on the general shape of the central figure, and at the end of the speech paragraph the description concerns parts (arms and palms) of it.

3RD SPEECH PARAGRAPH: TOPICAL TRANSITION TO DETAILS

12.2 st

07 ↑VAsemmalta Olalta ROIKkuu (.)
left-ABL shoulder-ABL hangs
PEHmeästi LASkostuva YÖNsininen viitta (0.3)
softly folding night-blue cloak
joka Ulottuu MAAhan asti. (2.5)
that reaches ground-ILL POSTP
On the left shoulder hangs a night-blue cloak
that reaches the ground.

The third speech paragraph consists of one spoken sentence (line 07). It starts 7.2 semitones higher than the preceding spoken sentence. Thus, the change of paragraph is very clear. There is a transition in the description of details from the physical composition to other objects. Thematically, the point of view moves slightly away from the central figure: in the preceding speech paragraph, the AD focused on the general shape of the central figure, and now it concerns an element (the cloak) which is not really a part of the central figure but a garment hanging on it.

5 Discussion

Previous studies have already shown that a sentence-initial high pitch acts as a discourse-structuring device in simultaneous interpreting and in speech-to-text interpreting (Nafá Waasaf 2007; Wiklund 2014). Our analysis of two audio describers' pitch and segmentation of information in a corpus of audio described art in Finnish indicates that the same applies to audio description. According to our study, the recorded AD speech consists of speech paragraphs, the beginnings of which are marked with a sentence-initial high pitch. Inside the same paragraph, each spoken sentence typically starts on a lower pitch level than the preceding one. This "supradeclination" can, however, be violated for example when the speaker emphasizes a word. As Wichmann (2000: 121) observes, the supradeclination is not systematic in data coming from naturally occurring situations, as in our case.

In addition, our study suggests that there is a relationship between a rise of pitch level and a topical transition in audio description. When the topical transition is clear (for example, when the topic of description shifts from one entity to another), the rise of pitch between the beginnings of two consecutive spoken sentences is large (see, for instance, Example 1, 1st

speech paragraph). Furthermore, the prosodic presentation – here, variation in the sentence-initial pitch – can cue topical transitions in the descriptions and, ultimately, changes in the ways of perceiving visual art: for instance, when a shift in attention occurs from contents to style or from the general impression to details.

This symmetry between the transition in prosody and the transition in information is, however, not always systematic, for instance when a new speech paragraph begins but the theme of the description remains the same (as in Example 3, between the 2nd and 3rd speech paragraphs). Though here as well, it is possible to interpret the topic of the new speech paragraph as a new entity (for example, the shift from a general introduction to a detailed description, as in Example 3). There is enough evidence in the research literature that a pitch reset indicates a shift of some kind. Accordingly, if a sentence-initial high pitch is produced in a place where there is no shift of any kind, either it is an interesting case of interpretation, and thus justifiable, or it is poor practice and likely to lead to difficulties in processing for the listener.

Our findings also suggest that prosodic stress is typically strongly expressed in audio describers' speech: as the transcriptions of the examples show, both speakers frequently produce syllables that carry a saliently strong stress. This feature also seems to contribute to the discourse structuring and to the segmentation of information in audio descriptions. As the current study has, however, been focused on the role of the sentence-initial high pitch, further studies are needed to confirm the role of stress in this type of data.

6 Conclusions

Considering that audio description is a verbal-vocal activity where text is the intermediate and speech the final stage of translating visual images to hearing ears, it is striking how little research we find thus far on the interplay of text and voice in AD. The present study attempts to address this lack. As phenomena related to so called 'paragraph intonation' have not been studied much in Finnish data before (see however Aho 2010; Lehtinen 2010, Wiklund 2014, 2018), this study provides more evidence about the existence of the paragraph intonation in monologous Finnish speech. This, in turn, proves that although the Finnish intonation system (Iivonen 1998) is, generally speaking, very different from that of English (see, for instance, Bolinger 1998; Hirst 1998), concerning the discourse structuring functions of prosody, there are some significant similarities as well (Wichmann 2000). Thus, even if this study is based only on Finnish data, the findings can be expected to apply at least in a certain measure to many other languages. On the other hand, our database is far too small (the number of informants = 2) for us to draw any gender-based conclusions.

In this paper, we have treated the prosodic realization of AD as an indication to the describers' cognitive processing, that is, their visual perception and cognitive structuring of information. Nonetheless, the prosodic realization is part of a vocal performance which in turn is influenced by other factors as well, such as emotions and the general physiological state (Liebenthal *et al.* 2016; Väyrynen 2014; Wilson & Wharton 2006). When, for instance, the AD makes an asymmetrical prosodic shift, the reason may be found in the performance: the describer may get tired if the AD continues for very long or the concentration may get disrupted. These effects on the performance can cause inappropriate prosodic cueing, such as new beginnings at illogical points. Finally, it is an individual matter when describers perceive a change of topic to be significant enough for prosodic cueing. It is also noteworthy that, according to Wilson

and Wharton (2006: 1560), "[…] the effects of prosody may be either accidental or intentional, and if intentional, either covertly or overtly so."

The cues of prosody and the power of voice overall are important issues to be brought in to the AD training. It is a well-known fact among AD professionals and users that reflecting a holistic visual representation by the linear means of language is a problem. This means that people listening to AD speech often get "lost in the picture" the longer the description lasts and the more details that are given. To cope with the problem, instructions are given that the AD should consist of neat sentences, providing one information per sentence. One solution to this problem can be the systematic use of prosodic cues, such as the sentence-initial high pitch and the organization of the AD into speech paragraphs. Thus, the speech paragraphs are not only prosodic and topical wholes but also visual wholes.

The hybrid form of AD being both text and speech has consequences for its use and, simultaneously, the role of prosodic cueing. In contrast to spoken delivery, AD can also be offered as written text,[12] which enables the use of the text according to personal preferences; for instance, to have it read out loud by a speech synthesis in text-to-speech software (TTS) that visually-impaired users tend to apply for reading. However, users seem to prefer the human voice (Szarkowska and Jankowska 2012, Fernandéz-Torné and Matamala 2015). The automatic TTS conversion is not able to produce social prosody, i.e. "the subtle differences in intonation and voice quality [that] are social signals", in the way humans do (Campbell and Li 2015: 99–100). On the other hand, the text format is useful because it enables a more independent contemplation of the artwork: the users can go through the description in the speed

[12] For an example, see the Verbal Description Database at the Art Beyond Sight Institute: http://www.artbeyondsight.org/mei/verbal-description-training/samples-of-verbal-description/ (6.2.2018)

and order they prefer. The TTS software is able to signal paragraph change and structural aspects if these are marked typographically. The synthetic voice may also be perceived as neutral – non-human – and thus corroborate the feeling of objectivity that is traditionally being considered important in audio description.[13]

All in all, how the voicing affects the use or enjoyment of AD is a subject for further research: for instance, are descriptions that are rendered "freely" in speech and have properties of spoken rather than written language cognitively less demanding to follow? In the future, it is likely that the human voice and the TTS are used for different purposes and in different communicative situations: the former for immersing people in artistic experiences and the latter for more mundane tasks, such as information search in digital content. Indeed, as museums and archives are going online, the potential of AD multiplies. The online format of art when combined with AD makes it easier to access the art works from home and other places without the need of traveling to the museum. Moreover, the more the visual and audiovisual contents are being verbally and textually described, the more readily they are found on the internet via text-based search engines.

Acknowledgements

The authors thank Academy of Finland (the research project Multimodal Translation with the Blind, grant number 295104) and Helsinki Collegium for Advanced Studies for the financial and scientific support of this research. We are also grateful to Ateneum and Sara Hildén Art Museum for providing us with the valuable research data.

[13] Recent research indicates that alternative, more subjective approaches to producing AD may be more apt when engaging visually impaired audiences more profoundly in an artistic experience (e.g. Walczak & Fryer 2017).

Appendix

Transcription symbols

(0.4)	A pause and its duration (seconds)
(.)	A micropause (less than 0.2 seconds)
.	Falling intonation
;	Slightly falling intonation
,	Continuing intonation
?	Rising intonation
¿	Slightly rising intonation
↑	METsäaukealla Raised pitch level
TUIjottavat	Stressed syllable
.hhh	Inbreath

References

Aho, Eija. 2010. *Spontaanin puheen prosodinen jaksottelu* [Prosodic segmentation of spontaneous speech]. Helsinki: University of Helsinki dissertation. http://urn.fi/URN:ISBN:978-952-10-6405-0 (accessed 21 July 2020).

Arminen, Ilkka. 2016. *Institutional interaction: Studies of talk at work.* New York: Routledge. http://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=nlebk&AN=1480500&site=ehost-live&scope=site (accessed 21 July 2020).

Ateneum. 2019. Kuvailutulkkaukset [Audio descriptions]. https://ateneum.fi/opastukset/kuvailutulkkaukset/# (accessed 21 July 2020).

Boersma, Paul & David Weenink. 2017. *Praat: Doing phonetics by computer* [Computer program]. Version 6.0.27. Retrieved 17 March 2017 from http://www.praat.org/

Bolinger, Dwight. 1998. Intonation in American English. In Daniel Hirst & Albert Di Cristo (eds.), *Intonation systems. A survey of twenty languages*, 45–55. Cambridge: Cambridge University Press.

Campbell, Nick & Ya Li. 2015. Expressivity in interactive speech synthesis; some paralinguistic and nonlinguistic issues of speech prosody for conversational dialogue systems. In Keikichi Hirose & Jianhua Tao (eds.), *Speech prosody in speech synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, 97–107. Berlin: Springer.

Chafe, Wallace L. 1980. The deployment of consciousness in the production of narrative. In Wallace Chafe (ed.), *The Pear Stories: cognitive, cultural and linguistic aspects of narrative production*, 9–50. Norwood, NJ: Ablex.

Chafe, Wallace L. 1994. Discourse, consciousness and time: The flow and conscious experience in writing and speaking. Chicago: The University of Chicago Press.

Couper-Kuhlen, Elizabeth. 1986. *An introduction to English prosody.* Tübingen/London: Niemeyer/Arnold.

Couper-Kuhlen, Elizabeth. 2000. Prosody. In Jef Verschueren, Jan-Ola Östman, Jan Blommaert & Chris Bulcaen (eds.), *Handbook of Pragmatics*, 1–19. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/hop (accessed 21 July 2020).

Couper-Kuhlen, Elizabeth. 2006. Prosodic cues of discourse units. In Keith Brown (ed.), *Encyclopedia of Language & Linguistics*, 2nd edn. 178–182. http://dx.doi.org/10.1016/B0-08-044854-2/00588-5 (accessed 21 July 2020).

Crystal, David. 1969. *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Crystal, David. 1980. *A first dictionary of linguistics and phonetics*. London: Deutsch.

De Coster, Karin & Volkmar Mühleis. 2007. Intersensorial translation. Visual art made up by words. In Jorge Diaz Cintas, Pilar Orero & Aline Remael (eds.*), Media for All: Subtitling for the deaf, audio description and sign language*, 189–200. Amsterdam: Rodopi.

Fernandéz-Torné, Anna & Anna Matamala. 2015. Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan. *JosTrans* 24. 61–88. https://www.jostrans.org/issue24/art_fernandez.pdf (accessed 21 July 2020).

Fix, Ulla (ed.). 2005. *Hörfilm: Bildkompensation durch Sprache*. Berlin: Erich Schmidt.

Fresno, Nazaret. 2014. Is a picture worth a thousand words? The role of memory in audio description. *Across Langauges and Cultures* 15(1). 111–129. https://akjournals.com/view/journals/084/15/1/article-p111.xml (accessed 21 July 2020).

Fryer, Louise. 2016. *An introduction to audio description. A practical guide.* London: Routledge.

Gutenberg, Norbert (2000). Mündlich realisierte schriftkonstituierte Textsorten (mrskT). In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager (eds.), *Text- und Gesprächslinguistik / Linguistics of Text and Conversation (Halbbd. 1/Vol. 1)*, 574–582. Berlin: Gruyter. http://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=nlebk&AN=186385&site=ehost-live&scope=site (accessed 21 July 2020).

Hirst, Daniel. 1998. Intonation in British English. In Daniel Hirst & Albert Di Cristo (eds.), *Intonation systems. A survey of twenty languages*, 56–77. Cambridge: Cambridge University Press.

Hirvonen, Maija. 2012. Contrasting visual and verbal cueing of space: Strategies and devices in the audio description of film. *New Voices in Translation Studies* 8. 21–43.

Hirvonen, Maija. 2014. *Multimodal representation and intermodal similarity: Cues of space in the audio description of film.* Helsinki: University of Helsinki dissertation. http://urn.fi/URN:ISBN:978-951-51-0369-7 (accessed 21 July 2020).

Iglesias-Fernández, Emilia, Silvia Martínez-Martínez & Antonio Javier Chica Núñez. 2015. Cross-fertilization between reception studies in audio description and interpreting quality assessment: The role of the describer's voice. In Jorge Díaz-Cintas & Rocío Piñero-Baños (eds.), *Audiovisual Translation in a Global Context*, 72–94. London: Palgrave Macmillan.

Iivonen, Antti. 1998. Intonation in Finnish. In Daniel Hirst & Albert Di Cristo (eds.), *Intonation systems. A survey of twenty languages*, 311–327. Cambridge: Cambridge University Press.

Kluckhohn, Kim. 2005. Informationsstrukturierung als Kompensationsstrategie – Audiodeskription und Syntax. In Ulla Fix (ed.), *Hörfilm: Bildkompensation durch Sprache*, 49–65. Berlin: Erich Schmidt.

Koskela, Anna. 2013. Aikuisten puhe- ja artikulaationopeus sekä artikulaationopeuden yhteys oraalimotorisiin taitoihin [Adults' speech and articulation rates and the connection between the articulation rate and oral-motor skills]. Oulu: University of Oulu MA thesis. http://urn.fi/URN:NBN:fi:oulu-201312102031 (accessed 18 February 2019).

Kreiman, Jody. 1982. Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics* 10(2). 163–175. https://www-sciencedirect-com.libproxy.tuni.fi/journal/journal-of-phonetics/vol/10/issue/2 (accessed 21 July 2020).

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the representation of mental referents in discourse.* Cambridge: Cambridge University Press.

Laver, John. 1994. *Principles of phonetics.* Cambridge: Cambridge University Press.

Lehtinen, Mari. 2010. The recategorisation of the rheme and the structure of the oral paragraph in French and in Finnish. *Discours* 7. https://doi.org/10.4000/discours.8007 (accessed 21 July 2020).

Liebenthal, Einat, David A. Silbersweig & Emily Stern. 2016. The language, tone and prosody of emotions: Neural substrates and dynamics of spoken-word emotion perception. *Frontiers of Neuroscience* 10(506). https://doi.org/10.3389/fnins.2016.00506 (accessed 2 April 2019).

Maszerowska, Anna, Anna Matamala & Pilar Orero (eds.). 2014. *Audio description. New perspectives illustrated.* Amsterdam: John Benjamins. http://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=e000xww&AN=868017&site=ehost-live&scope=site (accessed 21 July 2020).

Mazur, Iwona & Jan-Louis Kruger. 2012. Pear Stories and audio description: Language, perception and cognition across cultures. Special issue of *Perspectives: Studies in Translation Theory and Practice* 20(1). https://www.tandfonline.com/toc/rmps20/20/1 (accessed 21 July 2020).

Nafá Waasaf, María Lourdes. 2007. Intonation and the structural organisation of texts in simultaneous interpreting. *Interpreting* 9(2). 177–198.

Neves, Josélia. 2012. Multi-sensory approaches to (audio) describing visual art. *MonTi* 4. 277–293. https://doi.org/10.6035/MonTI.2012.4.12 (accessed 2 April 2019).

Poethe, Hannelore. 2005. Audiodeskription – Entstehung und Wesen einer Textsorte. In Ulla Fix (ed.), *Hörfilm: Bildkompensation durch Sprache*, 33–48. Berlin: Erich Schmidt.

Ramos, Marina. 2015. The emotional experience of films: Does audio description make a difference? *The Translator* 21(1). 68–94. https://doi.org/10.1080/13556509.2014.994853 (accessed 21 July 2020).

Remael, Aline, Nina Reviers & Gert Vercauteren (eds.). 2015. Pictures painted in words: ADLAB Audio Description guidelines. Trieste: Edizioni Università di Trieste. http://hdl.handle.net/10077/11838 (accessed 21 July 2020).

Sluijter, Agaath & Jacques Terken. 1993. Beyond sentence prosody: paragraph intonation in Dutch. *Phonetica* 50. 180–188.

Snyder, Joel. 2008. Audio description: The visual made verbal. In Jorge Díaz Cintas (ed.), *The didactics of audiovisual translation*, 191–198. Amsterdam/Philadelphia: John Benjamins. http://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=e000xww&AN=243195&site=ehost-live&scope=site (accessed 21 July 2020).

Soler Gallego, Silvia. 2018a. Audio descriptive guides in art museums. A corpus-based semantic analysis. *Translation and Interpreting Studies* 13(2). 230–249. https://doi.org/10.1075/tis.00013.sol (accessed 21 July 2020).

Soler Gallego, Silvia. 2018b. Intermodal coherence in audio descriptive guided tours for art museums. *Parallèles* 30(2). 111–128. https://www.paralleles.unige.ch/fr/tous-les-numeros/numero-30-2/ (accessed 21 July 2020).

Szarkowska, Agnieska & Anna Jankowska. 2012. Text-to-speech audio description for voiced-over films. A case study of audio described Volver in Polish. In Elisa Perego (ed.), *Emerging topics in translation: Audio description*, 81–98. Trieste: Edizione Universita di Trieste. http://hdl.handle.net/10077/6356 (accessed 21 July 2020).

Väyrynen, Eero. 2014. *Emotion recognition from speech using prosodic features.* Oulu: University of Oulu dissertation. http://jultika.oulu.fi/files/isbn9789526204048.pdf (accessed 21 July 2020).

Walczak, Agnieszka & Louise Fryer. 2017. Creative description: The impact of audio description style on presence in visually impaired audiences. *British Journal of*

Visual Impairment 35(1). 6–17. https://doi.org/10.1177/0264619616661603 (accessed 21 July 2020).

Wichmann, Anne. 2000. *Intonation in text and discourse: Beginnings, middles and ends.* Harlow: Pearson Education Limited.

Wiklund, Mari. 2014. The realization of pitch reset in Finnish print interpreting data. *Text & Talk* 34(4). 491-520. https://www.degruyter.com/view/journals/text/34/4/article-p491.xml (accessed 21 July 2020).

Wiklund, Mari. 2018. Indicating dependency between spoken sentences by prosodic means. *Discours* 22 (https://doi.org/10.4000/discours.9675) (accessed 21 July 2020).

Wilson, Deirdre & Tim Wharton. 2006. Relevance and prosody. *Journal of Pragmatics* 38(10). 1559–1579. https://doi.org/10.1016/j.pragma.2005.04.012 (accessed 21 July 2020).