Research Article

Lorena Bort-Mir and Marianna Bolognesi*

Reliability in the identification of metaphors in (filmic) multimodal communication

https://doi.org/10.1515/mc-2021-0022
Received November 24, 2021; accepted June 9, 2022; published online July 19, 2022

Abstract: Research on multimodal communication is complex because multimodal analyses require methods and procedures that offer the possibility of disentangling the layers of meaning conveyed through different channels. We hereby propose an empirical evaluation of the Filmic Metaphor Identification Procedure (FILMIP, Bort-Mir, L. (2019). *Developing, applying and testing FILMIP: the filmic metaphor identification procedure*, Ph.D. dissertation. Universitat Jaume I, Castellón.), a structural method for the identification of metaphorical elements in (filmic) multimodal materials. The paper comprises two studies: (i) A content analysis conducted by independent coders, in which the reliability of FILMIP is assessed. Here, two TV commercials were shown to 21 Spanish participants for later analysis with the use of FILMIP under two questionnaires. (ii) A qualitative analysis based on a percentage agreement index to check agreement among the 21 participants about the metaphorically marked filmic components identified on the basis of FILMIP's seven steps. The results of the two studies show that FILMIP is a valid and reliable tool for the identification of metaphorical elements in (filmic) multimodal materials. The empirical findings are discussed in relation to multimodal communication open challenges.

Keywords: FILMIP; metaphor identification; multimodal communication; multimodal metaphors

1 Introduction

One of the major challenges in the analysis of multimodal communication is the development of procedures and protocols that can be applied by different scholars in the same way. This is needed to ensure that the differences observed across data sets can be interpreted as genuine differences, rather than as noise due to the fact that different analysts applied the same procedure in different ways.

Generally speaking, when human judgment is involved in the analysis and classification of any type of data, unwanted variability is usually very high (Kahneman et al. 2021). Such variability can be attributed to different sources of error. In front of the same judging task, different analysts can behave in different ways on different occasions, for distinct reasons: some judges are inherently more severe than others, some judges may be influenced by external factors that are irrelevant to other judges, and so forth. The analysis of multimodal communication is particularly difficult, because of the variety of information encoded, inferred, and entangled within multimodal messages. To limit such unwanted variability, it is important that the tools provided to the analysts to perform their judgment tasks are well designed. In this way, it is more likely that independent analysts may approach the task in comparable ways. When this is the case, the method used for

***Corresponding author: Marianna Bolognesi**, Dipartimento di Lingue, University of Bologna, Letterature e Culture Moderne, via Cartoleria 5, Bologna, Emilia-Romagna, 40100, Italy, E-mail: m.bolognesi@unibo.it. https://orcid.org/0000-0002-3292-8968
Lorena Bort-Mir, Lingüística Aplicada, Universitat Politècnica de València, Valencia, Valencia, Spain, E-mail: lbormir@upv.es. https://orcid.org/0000-0003-0067-6492

the analysis is considered to be reliable and it stands a higher chance to deliver the same results when the study is replicated by different analysts. Reliability, as defined by Krippendorff (2011: 94), "is the extent to which different methods, research results, or people arrive at the same interpretations or facts".

This paper investigates the reliability of a method developed and used to analyze multimodal data, and in particular to identify whether a message expressed by filmic means contains metaphorical elements. This is a non-trivial task because it adds to the complexity given by multimodality, the identification of metaphorical elements, where metaphors themselves are difficult to define and identify. The Filmic Metaphor Identification Procedure (FILMIP; Bort-Mir 2019) is a procedure developed for the identification of filmic elements that are used metaphorically, in multimodal (filmic) communication. We hereby show how this procedure delivers reliable and thus arguably replicable analyses of multimodal (filmic) communications. We present two empirical studies in which FILMIP was applied by independent analysts for the identification of metaphors in filmic advertisements. We also discuss critical aspects and potential limitations of the method under scrutiny, as well as the broader implications for multimodal communication.

2 Theoretical background

The scientific literature on content analysis of multimodal communication shows great variability in the approaches adopted by different analysts, in different disciplines. The tasks are varied and not easily comparable. Content analyses of multimodal communication may have different goals and address different research questions. They may involve the classification of users' configurations of emoji into communicative strategies used to express different types of concepts (Danesi 2016; Das et al. 2019; Wicke and Bolognesi 2019). They may involve classifications of sign language descriptions of spatial relations between objects shown in pictures (Ortega et al. 2017), or they may involve classifications of gestures to investigate their different possible communicative functions (Cienki 2016; Kopp et al. 2008; Mittelberg 2019).

Even though there are some attempts to automate the process of metaphor detection and interpretation in written language (Fass 1991; Mason 2004; Su et al. 2017; Wilks 1978, among many other studies), the identification of metaphors in multimodal communication is considered complex phenomena that cannot, for the moment, be identified automatically. Thus, content analyses performed by human judges are commonly used in fields such as cognitive linguistics, to identify and manually classify metaphorical constructions in texts, visuals, or audiovisuals. The goals of these studies may be diverse, ranging from the classification of different types of pictures into metaphorical and non-metaphorical ones (e.g., Bolognesi et al. 2017; Stampoulidis and Bolognesi 2019), or the classification of different types of metaphors and other figurative constructions in multimodal advertisements (Kjeldsen and Hess 2021; Pérez-Sobrino 2016). Conceptual Metaphor Theory (Lakoff and Johnson 1980) spread the research on metaphor in areas other than just linguistics, as is the case of cognitive neuroscience (Johnson 2010; Van Dijk et al. 2008) or psycholinguistics (Gibbs and Colston 2012), among many others. The main idea of this theory on the study of this particular trope is that metaphor is not just a linguistic device employed to embellish poetry or prose, but something that people use in their everyday life to make sense of abstract concepts that are difficult to explain or to talk about. The so famous ARGUMENT IS WAR conceptual metaphor is used through linguistic expressions such as "Your claims are indefensible", "I demolished his arguments" or "He shot down all of my arguments" (Lakoff and Johnson 1980:4). Thus, the way we structure the action that we perform when arguing is by means of one literal concept (war, the source domain, as it is the concept from which we use attributes that we map onto the other concept) being somehow compared to a more abstract concept (argument, the target domain, as it is the concept that we aim to talk about).

For the identification of metaphors in moving images, FILMIP has been recently proposed and it will be hereby evaluated. The procedure takes inspiration from related procedures, developed in the past decades, for the reliable identification of metaphors expressed in different semiotic systems. The pioneering procedure for metaphor identification, known as MIP (Metaphor Identification Procedure) was developed by the Pragglejaz Group (2007) and it focuses on linguistic metaphoric expressions used in written texts. MIP is intended to

lead analysts to the decision of whether a lexical unit is used metaphorically in a given context. It takes the researcher into the identification of the basic and also the contextual meaning of a lexical unit, both established under the definitions found in, at least, two dictionaries. The basic and contextual meanings must then be contrasted and compared, thus leading the analyst to the final decision of whether that lexical unit is metaphorically used. For instance, given a sentence like *Do you see what I mean?*, the word *see* would be considered an MRW (metaphor related word), as the contextual meaning from the sentence is found in the fourth entry in the dictionary as "to understand something" ("See", Cambridge English Dictionary, n.d.), and out of the Essential Meaning section in the Merriam-Webster as "to perceive the meaning or importance of" ("See", Merriam-Webster, n.d.). The basic meaning of the word is shown in the former dictionary as "to notice people and things with your eyes", and in the latter as "to notice or become aware of (someone or something) by using your eyes". Thus, as both basic and contextual meanings differ from each other, the MRW *see* could be marked for metaphoricity.

Relating certain linguistic forms to their underlying conceptual structures and determining which set of correspondences are involved in the correlation between two different cognitive domains (or cross-domain mappings) is one of the main issues tackled by Steen (2001, 2002, 2009) resulting in a new, refined version of MIP, called MIPVU (MIP plus the initials of the Vrije Universiteit, Steen et al. 2010). The difference between the two procedures is that whereas the MIP aims at identifying metaphorically used words in spoken and written discourse, MIPVU focuses on the identification of 'metaphor-related words', including here "all lexical units in the discourse that can be related to cross-domain mappings in conceptual structure" (Dorst 2011: 102). Thus, the MIPVU bases its metaphor identification on the well-established idea that if metaphor is a matter of thought, then the cross-domain mapping between domains can be understood as all phenomena that imply a connection of similarity between two distinct domains (Grady et al. 1999; Steen 2008; Zbikowski 1997).

Other procedures have been developed more recently for the identification of deliberate metaphors in language (DMIP, Reijnierse et al. 2018) and for the identification of other types of figurative language, notably hyperbole (HIP, Burgers et al. 2016) and verbal irony (VIP, Burgers et al. 2011). As for semiotic systems other than verbal language, a procedure for the identification of metaphor in still images has been developed and applied to various visual genres (VISMIP, Šorm and Steen 2013). This procedure focuses on the identification of visual units with a metaphorical use in different persuasive materials, such as political cartoons or advertisements. VISMIP differs from MIPVU in language in that the latter looks for differences between the basic and contextual meanings of lexical units within dictionaries; as there are no such dictionaries for images, VISMIP uses Wordnet Online, a big database of English words created by Princeton University where researchers can find nouns, adjectives, adverbs, and verbs classified and grouped according to their meanings. For instance, assuming that there is an ice-cream cone in a picture, but instead of the ice-cream scoop it is the Earth that is melting, we would search for *Earth* or *planet* in Wordnet and would compare the results with the search for the words *ice-cream scoop*. As the results would be different, it would be assumed that there is a comparison between both cognitive domains (the "planet Earth" domain versus the "melting ice-cream" domain), and the picture would be marked for metaphoricity.

FILMIP evolved from some basic theoretical assumptions that characterize the identification procedures described above. In particular, the seven steps of the procedure are analogous to MIPVU and VISMIP's steps: the three entail the identification of the basic and contextual meaning of the units under analysis, and the comparison between cognitive domains is based upon incongruity, for instance. However, due to the complexity of filmic pieces in comparison to language and still images, FILMIP diverges in several stages of the procedure from the other methods in order to adapt it to the intricacy of multimodal materials, as is the case of its first step, with both macro and micro analysis of the materials, where the videos should be decomposed to the smallest level of granularity in order to envisage some kind of abstract meaning and the message that each video is intended to express. For this to happen, the method deepens into what multimodal theory adds to metaphor research.

There is no consensus among the research community about the basic tenet for the construction of an appropriate theory on multimodal metaphor, even though there is remarkable research done in the field (Forceville and Uriós-Aparisi 2009; Jewitt 2011; Kappelhoff and Müller 2011; Müller and Cienki 2009;

Rossolatos 2013, 2014). Some authors (Cienki and Müller 2008; Müller and Cienki 2009) claim that multimodal metaphors are constructed in various articulatory modalities, which entail, in the context of film theory, all the tools that cinema has to express meaning, that is, visuals, sounds, editing techniques, and the like. Forceville and Urios-Aparisi (2009) supports this claim and adheres to it in what he defines as a multimodal metaphor. According to the author, a multimodal metaphor is a metaphor in which the target and the source domains are represented in different communicative acts, semiotic resources (Pun 2008), perception channels (Norris 2013), or "modes" (Forceville and Urio-Apaisi 2009). These modes are mainly spoken and written language, visuals, sound, gestures, smell, taste, and touch. As an example, the linguistic metaphor PLANET WARMING IS ICE-CREAM MELTING could be depicted in a picture with a slogan or motto and planet earth melting like ice-cream. In this case, we have the visuals plus the written language, both domains of the metaphor use different communicative modes to be expressed (see Figure 1).

Regarding multimodal metaphors in films, FILMIP endeavors the complex task of identifying these tropes in audiovisual materials (see Figure 2 for the visualization of the procedure as a whole). The method entails a set of seven steps within two different phases. In Phase 1, researchers go into an extensive analysis of the multimodal content, describing the general meaning of the materials by the identification and description of all their communicative modes, for later identification of a plausible abstract meaning and message. In Phase 2, the analyst submerges into the concrete identification of metaphorical components in a similar fashion as in VISMIP.

3 Method

The two studies are briefly introduced as follows:

1. Study 1: content analysis intended to test the reliability of the application of the seven steps. It entails a content analysis based on the Kappa and Alpha reliability coefficients to check agreement among the coding scheme developed in order to carry out study 2 (look into metaphor identification). Two different TV commercials were projected to 21 participants who had to analyze those commercials with the use of

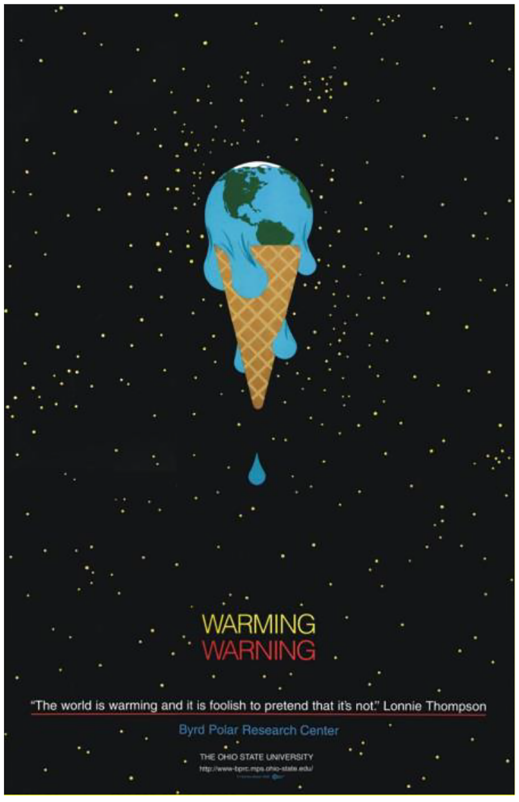

Figure 1: Global warming poster designed by the Byrd Polar Research Center, Ohio State University.

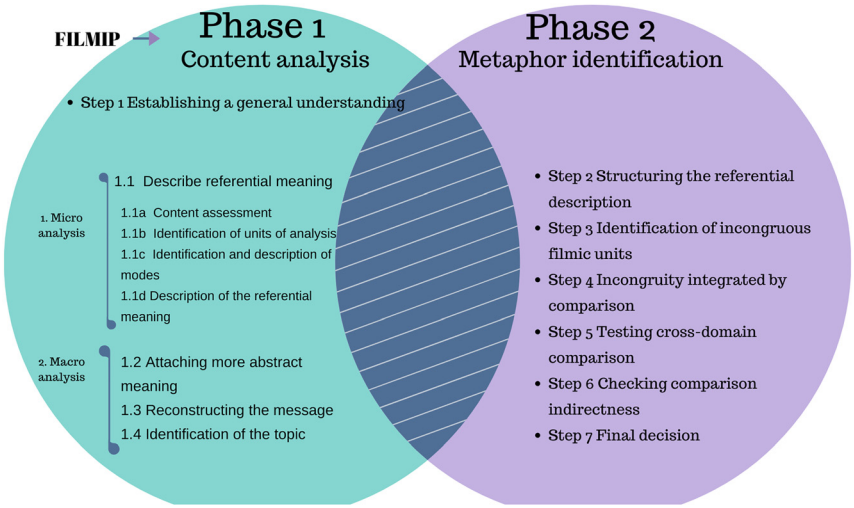

Figure 2: FILMIP's seven steps.

FILMIP. They had to fill out two questionnaires containing questions related to the procedure's seven steps. The content analysis was carried out in order to assess and classify the content of what the participants wrote about the filmic ads. The results obtained from Study 1 respond to the question of whether the annotators (three coders) classified the content in similar ways and to what extent these independent annotators gave similar classifications about the meaning encoded in the textual data from the questionnaires.

2. Study 2: a qualitative analysis based on a percentage agreement index to check agreement among analysts about the metaphorically marked filmic components identified on the basis of FILMIP's seven steps by responding to the following question: does FILMIP lead to identifying the same metaphorical elements by all analysts? Figure 3 summarizes the design of the empirical studies reported and discussed in the current paper.

3.1 Materials

For the two studies (Study 1 and Study 2) two different TV commercials from perfume brands were used: *Agua Fresca de Rosas* (Adolfo Domínguez 2015) and *Black Opium* (Yves Saint Laurent 2015). Links can be found in the online repository. The *Agua Fresca de Rosas* commercial is about a lady on a boat fishing roses in the sea with a net. The *Black Opium* commercial presents the perfume as something that a lady is searching for desperately.

Two online questionnaires[1] were designed for comparability results. The two questionnaires were distributed to participants after a 2-h training session on FILMIP, and they contained structured questions that corresponded to the most 'metaphor-clarifying' steps of the procedure. The questions were the following:

1. Which is the message of the commercial? This question corresponds to FILMIP's step 1.3 (reconstructing the message) within step 1 (establishing a general understanding). We assume that viewers are able to catch the message of the TV commercial only when a general understanding of it has been achieved.
2. Which incongruous units do you see, in case there are any? This question corresponds to step 3, identification of incongruous filmic units. Once the viewer understands the commercial and its message, he/she is supposed to see if there is anything in the commercial that is strange, alien, or weird, compared to its overall topic.

1 Data was collected in compliance with the General Data Protection Regulation 2016/679 of the European Parliament and of the Council of 06 April 2016. All the materials related to this study are stored on Open Science Framework at the following link: https://osf.io/3hq65/?view_only=f3c6e10a9d4b4ec6a600a62424d66c31.

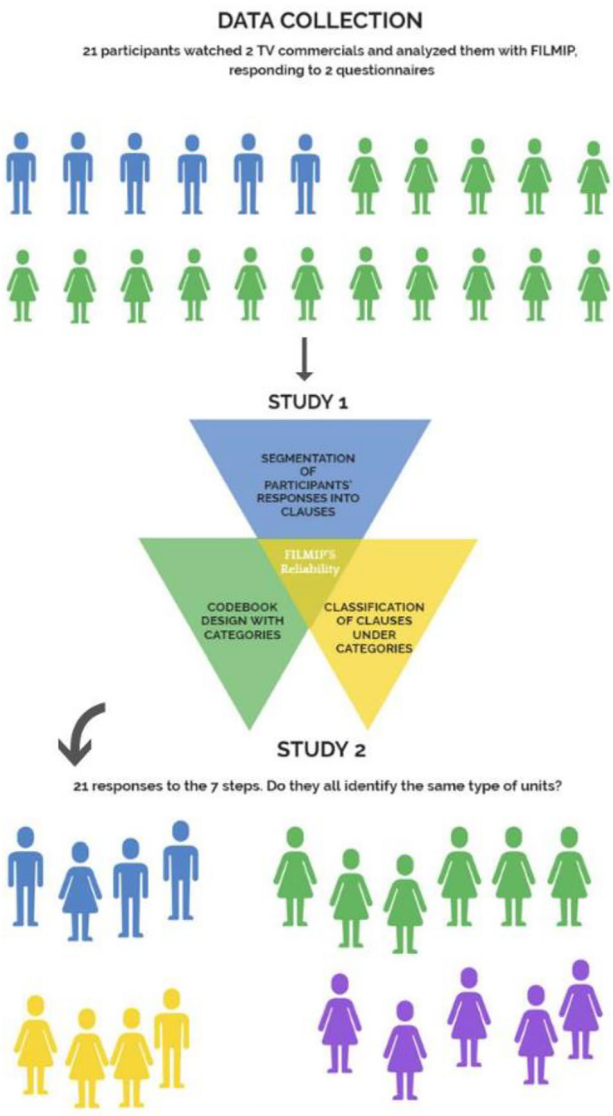

Figure 3: Design of the empirical studies.

3. Which is the comparison that you see in case there is any? This question corresponds to step 4, in which analysts must detect if that strange thing identified in the previous step is being depicted in the TV commercial as a comparison to another thing. If this is so, then the analyst would come up with the two domains of the metaphor (following our example of global warming, this comparison would be planet Earth and ice-cream).
4. Can you describe the metaphor if there is any? (A IS B) This fourth question corresponds to FILMIP's step 6, as at this stage, the viewer is supposed to "get" why these two different things from question 3 are being compared.
5. Is there a metaphor? (Corresponding to step 7).

The questionnaires were administered online through Google Forms[2] on an anonymous basis.

2 Ethical standards were accepted by the hosting institution of the first author, in accordance with Article six of the Regulations of the Deontological Commission and the Animal Welfare Ethics Committee of the Universitat Jaume I.

3.2 Participants

Participants[3] for Study 1 (content analysis, undertaken in 2018) were three coders engaged in the annotation process: (i) the project leader (trained annotator), female and aged 38, from the Doctoral Program Applied Languages, Literature and Translation at Universitat Jaume I, Castellón (Spain), (ii) one collaborator (trained annotator), male and aged 38, professor at Universitat Jaume I from the English Studies Department, and (iii) a novice coder, female and aged 30, associate professor at Universitat Jaume I, also from the English Studies Department.

Participants[4] for the data collection were students from the English Semantics course at Universitat Jaume I (UJI), Castellón (Spain). They were in their third year of the English Studies Degree. Participants were aged between 20 and 27. All of them were from Spain. The first questionnaire was filled in by 15 female and six male participants. The second questionnaire was completed in another session by 15 female and three male participants. All of them signed a consent form for their participation in the study. That form contained the required information about the implications of the study and the instructions for them to follow, not including specific details about the task. Complete details about the purposes of the questionnaires were given at the end of the course to those students who asked for them.

3.3 Procedure

Two different procedures were adopted depending on the study to be performed:
1) Procedure for data collection:

Data was collected in December 2017. Participants took a 2-h training session on FILMIP in class during their English Semantics course, then they individually analyzed the two TV commercials and filled in the forms in the other two consecutive 2-h sessions. Participants received some guidelines as to how to proceed with their analyses. The first TV commercial was projected five times in class, and then the students were asked to apply FILMIP individually (they could use their notebooks for their analyses). The participants were requested to fill in the first online form as they finished their analyses of the ad. Twenty-one participants completed this first task.

The second TV commercial was projected five times again in another 2-h session in class, and the students followed the same instructions (individual application of FILMIP and filling in the online form). Eighteen participants completed the second questionnaire.

As students did their analyses individually in class, a few doubts were solved about the procedural steps. However, they were not allowed to ask any questions or make any comments while performing their analyses about their results so as not to bias any of the responses from the other students or even from the instructors.

The participants' responses were collected online through Google Forms. All these individual responses (the data collected) from all the students were uploaded in two different spreadsheets (one spreadsheet per ad) to Google Docs. This data was segmented into clauses, as the clause is considered

> the basic semantic unit of discourse […] (and) the building blocks of text which are easiest to classify. When a group of words contained a subject and predicate of its own and formed part of a sentence, this was a clause. As a rule, the clauses corresponded to the basic units of discourse. (Šorm and Steen 2013: 16–17)

3 Participants for Study 1 are also referred to in the present work as coders and/or annotators, since they perform the two roles: they are the participants of the study and they code/annotate the data at the same time.

4 Participants for the data collection are also referred to in this thesis as *students* and/or *analysts*, since they can be included into these three roles: they are the participants of the studies, they are real students at university, and they are also the analysts of the commercials.

Participant ID	CLAUSE (segment)
1	With this perfume you can get any man
2	After analysing this advert, we can say that the girl of the advertisement is searching for a man (that she has to choose)
2	The man is represented by the flowers.
3	That the perfume is a tool that will help you to get new fresh men
4	With this perfume you can find love again.
5	That the perfume is a tool to find a new fresh man
6	With this perfume you will able to find the love again or at least another lover.
7	The message this commercial transmits is the sensuality that she feels while catching roses
7	and the aim of the commercial is to encourage the rest of the women to buy the perfume to make them feel in this way.
8	Men are roses, so roses are caught.
9	The perfume is used to attract men.
10	The message wants to sponsor a perfume
11	With that perfume is posible to find love
12	That you if you buy this perfume you will get a fragance like the roses in the commercial; fresh and soft.
13	The message is to create the desire of buying the parfum e
14	This fragance concentrates the power of a sea of roses in a tiny bottle.
15	The message of the commercial is that fresh water from roses are used for the perfume.
16	Hunting love
17	The woman is trying to get a man
18	The creator wants to transmit the message that with that perfume you can find the love
19	The fragance represents love
20	The message is that the women is pure as roses
21	The freshness of the perfume.

Figure 4: Segments extracted from questionnaire 1.

All the clauses or *segments* extracted from the participants' responses were separated into different sheets according to the five questions of the questionnaires (labeled under Question 1, Question 2, Question 3, Question 4, and Question 5. See Figure 4 below). The segments were in English, as the study was located within the English Studies Degree (at UJI).

2) Procedure for Study 1 (content analysis):

A coding book was elaborated according to the segments (the clauses produced by participants in the two written questionnaires), and a corresponding coding scheme was designed for each question. As some of the concepts within each segment were abstract, several categories were designed under the premise of inferring what the participants meant when they produced a particular thought (Barsalou 1992). The coding scheme is shown in the table below (Table 1):

The segments from Questionnaire 1 and Questionnaire 2 (about the two TV commercials) were coded with their corresponding category. Only one category was allowed for each segment. The categories were previously created and described according to the data obtained (the participants' responses).

Two codebooks were created in Google Spreadsheets with the following information:

- all the segments resulting from the data collected in the online questionnaires, separated into five sheets (one per each of the five questions of each questionnaire)
- annotation of segments according to the corresponding category by three independent coders.

A total of 114 segments were annotated for the analysis of the first questionnaire. A total of 100 segments were annotated for the analysis of the second questionnaire.

4 Analysis and results

4.1 Study 1: content analysis

Study 1 seeks to classify the written productions provided by the participants. Study 1 also forms the basis to perform Study 2 (the qualitative analysis where the agreement on the types of metaphors identified is tested in order to test the validity of the procedure).

Table 1: Coding scheme for the commercials designed according to the participants' responses.

Question	*Agua Fresca de Rosas*	*Black Opium*
Q.1: Which is the message of the commercial?	1. Buy perfume to get a man 2. Woman is purity 3. Men are roses 4. Feel sensual 5. Selling perfume	1. You need the perfume as you need drugs 2. Perfume makes you feel good as drugs do 3. Perfume is tool for women to be themselves 4. Selling perfume
Q.2: Which incongruous units have you found?	1. Nothing 2. Action – Object (FISH_FLOWERS) 3. Object – Location (ROSES_SEA) 4. Object (NET) 5. Object _ Action (SEA_SMELL) 6. Object (BOAT)	1. Nothing 2. Everything 3. Object_State (PERFUME_LIKE DRUG) 4. Music 5. Camera movements 6. Object _State (PUPIL-DILATED)
Q.3: What are they comparing?	1. FLOWERS(ROSES)_MEN (LOVE) 2. ROSES _ BEAUTY 3. ROSES_FISH 4. ROSES_SENSUALITY 5. WOMAN_GHOST 6. PERFUME_ROSE 7. NATURE_PERFUME 8. FISHING_SMELLING 9. ROSES_PURIFICATION 10. Irrelevant responses	1. PERFUME_DRUGS 2. SMELL OF PERFUME_ADDICTION 3. BIG PUPIL_SOMETHING SHE WANTS
Q.4: Do you think there is a metaphor in this commercial?	1. I don't know 2. Yes 3. No	1. Yes[a]
Q.5: Which is the metaphor that you see, if there is any?	1. Irrelevant responses 2. ROSES ARE MEN 3. MEN ARE FISH 4. FLOWERS ARE EROTISM 5. PERFUME IS SENSUALITY 6. PERFUME IS TOOL TO FIND LOVE 7. ROSE IS PERFUME 8. CATCHING ROSES IS FISHING 9. WOMAN IS ROSE	1. PERFUME BLACK OPIUM IS DRUG 2. PERFUME BLACK OPIUM IS ADDICTION

[a]The categories were created according to the participants' responses, and none of the participants responded with anything but a "yes" in this question from the *Black Opium* commercial.

Participants' responses to the questions in the two questionnaires about their analyses of the commercials were not on a yes/no basis. They explained their thoughts based on FILMIP's seven steps. Thus, in Study 1, the clauses (segments) resulting from the responses of the participants were classified into their corresponding categories (see Section 3.3) by building a coding scheme, and those annotations were investigated on a content analysis basis, tested by means of Krippendorff's Alpha[5] (2004, 2013) and Fleiss's Kappa (1971) coefficients,

5 Krippendorff defined reliability as the "extent to which different methods, research results, or people arrive at the same interpretations or facts" (2011: 94). FILMIP is tested under the Alpha index to investigate the extent to which the method leads different analysts to arrive at the same interpretations.

to check whether these categories reflect the real participants' responses. By evaluating the participants' responses, Study 1 evaluates the overall reliability of FILMIP.

The data analysis involved three phases:

4.1.1 Phase 1

Two independent and trained annotators coded all the participants' responses (classified into clauses) from the two questionnaires. The resulting coding scheme for the two questionnaires was quite simple containing the few categories described in Table 1 from the previous section. This simplicity implied no nested categories, and consequently, no discussion sessions for the revision and modification of the categories took place. For this reason, two coders were considered enough for Study 1 (content analysis).

The reliability test among these two first annotators was done with the online tool https://nlp-ml.io/jg/software/ira/#demo, "an online calculator for inter-rater agreement with multiple raters, featuring Light's Kappa, Fleiss's Kappa, Krippendorff's Alpha, and support for missing data" (Geertzen 2012).[6]

4.1.2 Phase 2

A training session of 1 h was performed with another annotator, the novice coder, as proposed in Bolognesi et al. (2017) in order to avoid previous mutual agreement and similar perspectives between the two first trained annotators (see also Krippendorf 2013). In this phase, the coding scheme was evaluated with the annotations of this new novice coder who was not aware of the aim of the study and who had never carried out a similar task before. The novice coder annotated the data relying only on the coding scheme. The educational background of the new non-trained coder was cognitive linguistics within a postgraduate course.

4.1.3 Phase 3

A formal reliability test was performed among all three coders to check agreement on all the categories developed for each of the questions from the two questionnaires. The same online tool mentioned above for Phase 2 was also used for this test (Geertzen 2012).

Given the characteristics of question 4 (*Is there a metaphor?*) resulting in a yes/no binary code, it was left out from the reliability test for this content analysis since Study 1 focuses on the type of segment written by each of the participants and not on the number of identified filmic metaphorical components.

The results from Phase 1 and Phase 3 are shown in the following table (Table 2).

4.2 Study 2: qualitative analysis

Study 2 leads to comparing whether the qualitative results of one analyst varies from those of other analysts. Thus, Study 2, based on a simple percentage agreement index[7] (Wu and Barsalou 2009), leads to check agreement about the type of metaphorically used filmic components identified by each analyst on the basis of FILMIP's seven steps. The results obtained offer a response to whether FILMIP leads independent analysts to identify the same type of metaphorical elements in films, hence checking the validity of the procedure.

The percentage agreement index for Study 2 was calculated with the aid of Google Docs Graphs in the same spreadsheet pertaining to the annotations.

---

6 This online tool is not available anymore, but the authors have found other similar tools that calculate the coefficients in a similar way: https://idostatistics.com/cohen-kappa-free-calculator/#calculator and https://www.graphpad.com/quickcalcs/kappa1/.

7 The field of study determines the acceptable agreement level. "If it's a sports competition, you might accept a 60% rater agreement to decide a winner. However, if you're looking at data from cancer specialists deciding on a course of treatment, you'll want a much higher agreement—above 90%. In general, above 75% is considered acceptable for most fields" (Glen, Lastly retrieved in June 2022).

Table 2: Reliability results from Phase 1 (two coders) and Phase 3 (three coders).

Question	TV commercial	Reliability test, two coders	Reliability test, three coders
Q.1: Which is the message of the commercial?	*Agua Fresca de Rosas*	$\kappa = 0.729$	$\kappa = 0.779$
		$\alpha = 0.73$	$\alpha = 0.78$
	Black Opium	$\kappa = 0.84$	$\kappa = 0.84$
		$\alpha = 0.84$	$\alpha = 0.84$
Q.2: Which incongruous units have you found?	*Agua Fresca de Rosas*	$\kappa = 1$	$\kappa = 0.88$
		$\alpha = 1$	$\alpha = 0.88$
	Black Opium	$\kappa = 0.71$	$\kappa = 0.90$
		$\alpha = 0.72$	$\alpha = 0.90$
Q.3: What are they comparing?	*Agua Fresca de Rosas*	$\kappa = 1$	$\kappa = 0.95$
		$\alpha = 1$	$\alpha = 0.96$
	Black Opium	$\kappa = 0.74$	$\kappa = 0.82$
		$\alpha = 0.75$	$\alpha = 0.83$
Q.5: Which is the metaphor that you see, if there is any?	*Agua Fresca de Rosas*	$\kappa = 0.94$	$\kappa = 0.96$
		$\alpha = 0.94$	$\alpha = 0.96$
	Black Opium	$\kappa = 1$	$\kappa = 1$
		$\alpha = 1$	$\alpha = 1$

With *Agua Fresca de Rosas* commercial (Adolfo Domínguez 2015), a total of 21 analyses (by 21 independent analysts, that is, the participants) per each of the five questions from Questionnaire 1 were taken into consideration for the calculation of percentages. With *Black Opium* commercial (Yves Saint Laurent 2015), a total of 18 analyses (by 18 independent analysts) per each of the five questions from Questionnaire 2 was investigated.

The results are shown below:

a) Percentages from the *Agua Fresca de Rosas* TV commercial:

The first percentage indexes were calculated on the questions from Questionnaire 1 with the following results:

- In question 1 regarding the message of the commercial, 52.5% of the participants agreed on the same message, "*you will find love if you buy the perfume*", 26.1% of the participants thought that the commercial was just explaining the properties of the perfume. A 4.3% identified the message as "*woman is purity*", and an 8.7% responded that the message was "*men are flowers*" and that "*the commercial wanted you to feel sensual*".
- In question 2 regarding the identification of incongruity, 50% saw that the location of the roses in the sea was strange; 15.4% thought that the action of fishing flowers was incongruous, and 11.5% responded that the net was the incongruous element in the commercial. Another 11.5% (3 participants) wrote irrelevant responses to the question itself. The fact that the sea was smelt was identified as incongruous by 7.7% of the participants, and another 3.8% (1 participant) thought that what was incongruous was the boat.
- In question 3 regarding the identification of comparison, 42.9% of the participants compared flowers with men, and 14.3% compared roses with fish. 9.5% saw a comparison between the perfume and roses. The rest of the comparisons obtaining a 4.8% each of them (1 participant) were between roses and beauty, roses and sensuality, woman and ghost, nature and perfume, fishing and smelling, and roses and purification.
- In question 4 regarding the detection of a metaphor, 43.5% of the participants identified the ROSES ARE MEN conceptual metaphor, and 13% identified the metaphor FISH ARE MEN. Four participants (17.4%) responded that they saw no metaphor. The rest of the conceptual metaphors were identified by one participant each (a 4.3% per each comparison), including FLOWERS ARE EROTISM, PERFUME IS SENSUALITY, PERFUME IS TOOL TO FIND LOVE, ROSE IS PERFUME, CATCHING ROSES IS FISHING, and WOMAN IS ROSE.

b) Percentages from *Black Opium* TV commercial:

The second percentage indexes were calculated about the questions from the commercial *Black Opium* (Yves Saint Laurent 2015):[8]

- In question 1 regarding the message of the commercial, 72% thought that it was "*you are addicted because you need the perfume*". 12% identified the message "*the perfume makes you feel good like drugs*", and another 12% identified it as "*the perfume is a tool for the woman daring to be herself thanks to drugs*". One of the participants (the remaining 4%) thought that the message was just selling perfume.
- In question 2 regarding the identification of incongruous units, 45% of the participants identified the perfume shown in the ad as a drug as the incongruity of the commercial. 20% thought that everything was incongruous, whereas another 10% saw no incongruity at all. Another 10% of the participants thought that the music was incongruous. Two of the participants (10%) identified the pupil dilating as the incongruous unit, and one remaining participant (5%) thought that the movements of the camera were the incongruous elements of the commercial.
- In question 3 regarding the identification of comparison, the results obtained are as follows: 73.7% of the participants compared the perfume with drugs, 21.2% compared the smell of the perfume with addiction, and 5.3% compared the eye and its big pupil with something that she wanted.
- Finally, in question 4 regarding the identification of a possible conceptual metaphor, 72.2% of the participants identified the conceptual metaphor PERFUME BLACK OPIUM IS DRUG, and the rest 27.8% identified the metaphor PERFUME BLACK OPIUM IS ADDICTION.

5 Discussion and conclusion

Two studies have been performed to test whether FILMIP can be considered a reliable tool for filmic metaphor identification. Study 1, the content analysis, led to the annotation of all the responses written by the participants in two questionnaires after applying FILMIP, according to a coding scheme that was tested for agreement by means of an analysis of the Kappa and Alpha indexes. The high indices obtained indicate that significant results[9] are achieved, with a mean value of 0.87 among two trained coders, and 0.89 among three annotators (trained plus novice).

The commercial *Agua Fresca de Rosas* (Adolfo Domínguez 2015), offers unanimous agreement about which incongruities are detected by the analysts and what is being compared in the ad. The final decision on whether there is a metaphor and what conceptual metaphor may be construed in the clip offers almost complete agreement. However, the result of the analysis of the first question (*which is the message of the commercial?*) is slightly lower (but still positive), with a 0.73 index. This may be due to the fact that question one is the vaguest of all, where participants had an ample range of thoughts for their responses. What this difference, though small, may be showing is that the more specific and concrete information we ask from the participants, the more unanimity can be reached with reliability tests.

However, as this difference does not exist in the analysis of the second commercial, we could state that a higher number of materials would lead to more refined conclusions about reliability metrics for multimodal analysis. More (and longer) materials with a higher number of elements to be analyzed would lead to more significant comparative results. The authors are also aware that genre may constitute an important factor in multimodal discourse analysis. In the future, the two studies conducted in this research will be applied to other

8 The video to which the analysis was performed is not available any more. It contained all director's cuts in which a pupil dilating could be observed, among many other elements that could be related to drug addiction.

9 As there is no consensus about the interpretation of coefficients, we take Landis and Koch (1977) Kappa values as valid for our study. Kappa values and strength of agreement: 0.8 (substantial agreement), 1.0 (perfect agreement). Regarding the Alpha values, we assume Krippendorff's (2013), requiring $\alpha \geq 0.800$ for any result to be acceptable, and a $\alpha \geq 0.667$ to be the lowest coefficient. This implies that a 0 result means perfect disagreement among coders, whereas getting a 1 means perfect agreement. Krippendorff indicates that all results equal or above 0.8 can be taken as very good results.

cinematic genres such as documentaries, cartoon films, feature films, and even music videos. Such analyses with more materials from different genres will then lead to a deeper understanding of the construct of multimodal communication, letting analysts know to what extent empirical research in this field can be attained for distinct purposes, implying that multimodal materials are good candidates for reliability tests.

Regarding the qualitative analysis (Study 2), it reports very high indexes as well. However, still, some differences among the responses of the participants (different percentage indexes) can be found, which may be due to several factors:

1. Multimodal metaphors in the filmic medium are complex tropes as they are construed cross-modally through a complicated composition of mappings between both domains of the metaphor (target and source), all depicted by the interaction and precise construction of the filmic tools that the filmmaker makes use of (Müller and Kappelhoff 2018). This entails that clustering a filmic multimodal metaphor in a simple A IS B formula is a difficult task (Forceville 2006, 2007) that depends on many variables (the context, the time, the analyst, the genre, etc.).
2. Inter-rater reliability based on percent agreement does not take agreement by chance into consideration. Further alternative methods to compute this agreement will be considered for future research.
3. Individual differences play a key role in the interpretation of metaphors, and cinema is not an exception. The cultural or social background of the analysts and even their level of expertise in metaphor research usually influence their perception and understanding of a given metaphor.

However, as FILMIP is not considered a tool for the analysis and interpretation of metaphors in filmic multimodal materials, but a tool to identify metaphorical elements in these materials to later perform a different kind of analysis, the authors consider that, as the studies offer quite significant percentages in the majority of the questions, we could state that the application of the Filmic Metaphor Identification Procedure allows analysts to perform replicable studies of multimodal structured-semiotic analysis of (filmic) multimodal materials by revealing all the existing levels of signification and their aesthetic configuration. The procedure, thus, not only entails a mere description of the filmic narrative but also the reconstruction of its meaning, aiming, during its first phase (the content analysis phase, including micro and macro analysis of the multimodal materials, see Figure 1 in Section 2), at decomposing and composing the communicative modes, arriving at the finest granularity possible of the multimodal artifacts under analysis.

Nonetheless, and in a second evaluative phase, we consider that it would be greatly insightful for the method's reliability to test the results obtained also by participants without the use of the procedure for comparability results.

To conclude, in a time where communication genres, media, and channels keep on expanding and growing, multimodal theory seems to be very well positioned for researchers in the field. Empirical methods that allow the scientific community to perform replicable multimodal analysis will raise the possibilities of multimodal communication to attain valuable global, interdisciplinary research.

References

Adolfo Domínguez [Eugenia Silva]. (2015). Adolfo Domínguez: Agua Fresca de Rosas [video file], Available at: https://www.youtube.com/watch?v=K2rjjhlloL8.

Barsalou, L. (1992). Frames, concepts, and conceptual fields. In: Lehrer, A. and Kittay, E.F. (Eds.), *Frames, fields, and contrasts*. Erlbaum, Hillsdale, NJ, pp. 21–74.

Bolognesi, M., Pilgram, R., and Van den Heerik, R. (2017). Reliability in content analysis: the case of semantic feature norms classification. Behav. Res. Methods 49: 1984–2001.

Bort-Mir, L. (2019). *Developing, applying and testing FILMIP: the filmic metaphor identification procedure*, Ph.D. dissertation. Universitat Jaume I, Castellón.

Burgers, C., Van Mulken, M., and Schellens, P.J. (2011). Finding irony: an introduction of the verbal irony procedure (VIP). Metaphor Symbol 26: 186–205.

Burgers, C., Brugman, B.C., Renardel de Lavalette, K.Y., and Steen, G.J. (2016). HIP: a method for linguistic hyperbole identification in discourse. Metaphor Symbol 31: 163–178.

Cienki, A. (2016). Cognitive Linguistics, gesture studies, and multimodal communication. Cognit. Ling. 27: 603–618.

Cienki, A. and Müller, C. (Eds.) (2008). *Metaphor and gesture*, Vol. 3. John Benjamins Publishing.
Danesi, M. (2016). *The semiotics of emoji: the rise of visual language in the age of the internet*. Bloomsbury Publishing, London.
Das, G., Wiener, H.J., and Kareklas, I. (2019). To emoji or not to emoji? Examining the influence of emoji on consumer reactions to advertising. J. Bus. Res. 96: 147–156.
Dorst, A.G. (2011). *Metaphor in fiction: language, thought and communication*, Doctoral thesis, Amsterdam.
Fass, D. (1991). met*: a method for discriminating metonymy and metaphor by computer. Comput. Ling. 17: 49–90.
Fleiss, J. (1971). Measuring nominal scale agreement among many raters. Psychol. Bull. 76: 378–382.
Forceville, C. (2006). Non-verbal and multimodal metaphor in a cognitivist framework: agendas for research. In: Kristiansen, G., Achard, M., Dirven, R., and Ruiz de Mendoza, F. (Eds.), *Cognitive linguistics: current applications and future perspectives*. Mouton de Gruyter, Berlin/New York, pp. 379–402.
Forceville, C. (2007). Multimodal metaphor in ten Dutch TV commercials. Public J. Semiot. 1: 15–34.
Forceville, C. and Urios-Aparisi, E. (Eds.) (2009). *Multimodal metaphor*. Mouton de Gruyter, Berlin.
Geertzen, J. (2012). Inter-rater agreement with multiple raters and variables. Available at: https://mlnl.net/jg/software/ira/ (Accessed 20 February 2015).
Gibbs, R.W. Jr. and Colston, H.L. (2012). *Interpreting figurative meaning*. Cambridge University Press, Cambridge.
Glen, Stephanie. (2022). "Inter-rater reliability IRR: definition, calculation" from StatisticsHowTo.com: elementary statistics for the rest of us! Available at: https://www.statisticshowto.com/inter-rater-reliability/ (Accessed 6 June 2022).
Grady, J., Oakley, T., and Coulson, S. (1999). Blending and metaphor. Amsterdam Stud. Theory Hist. Ling. Sci. Ser. 4: 101–124.
Group, P. (2007). MIP: a method for identifying metaphorically used words in discourse. Metaphor Symbol 22: 1–39.
Jewitt, C.E. (2011). *The Routledge handbook of multimodal analysis*. Routledge/Taylor & Francis Group, London.
Johnson, M. (2010). Metaphor and cognition. In: Schmicking, D. (Ed.), *Handbook of phenomenology and cognitive science*. Springer, Dordrecht, pp. 401–414.
Kahneman, D., Sibony, O., Fusaro, R., and Sperling-Magro, J. (2021 In this issue). Sounding the alarm on system noise. McKinsey Quarterly (3), Available at: https://www.mckinsey.com/quarterly/the-magazine/2021-issue-3-mckinsey-quarterly (Accessed 09 July 2022).
Kappelhoff, H. and Müller, C. (2011). Embodied meaning construction: multimodal metaphor and expressive movement in speech, gesture, and feature film. Metaphor Soc. World 1: 121–153.
Kjeldsen, J. and Hess, A. (2021). Experiencing multimodal rhetoric and argumentation in political advertisements: a study of how people respond to the rhetoric of multimodal communication. Vis. Commun. 20: 327–352.
Kopp, S., Bergmann, K., and Wachsmuth, I. (2008). Multimodal communication from multimodal thinking—towards an integrated model of speech and gesture production. Int. J. Semantic Comput. 2: 115–136.
Krippendorff, K. (2004). Reliability in content analysis: some common misconceptions and recommendations. Hum. Commun. Res. 30: 411–433.
Krippendorff, K. (2011). Agreement and information in the reliability of coding. Commun. Methods Meas. 5: 93–112.
Krippendorff, K. (2013). *Content analysis: an introduction to its methodology*, 3rd ed. Thousand Oaks, CA: Sage.
Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago.
Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics 33: 159–174.
Mason, Z.J. (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. Comput. Ling. 30: 23–44.
Müller, C. and Cienki, A. (2009). Words, gestures, and beyond: forms of multimodal metaphor in the use of spoken language. In: Charles Forceville and Eduardo Uriós-Aparisi (Eds.), *Multimodal metaphor*. Walter de Gruyter GmbH & Co KG, Berlin, pp. 297–328.
Müller, C. and Kappelhoff, H. (2018). *Cinematic metaphor: experience–affectivity–temporality*, Vol. 4. Walter de Gruyter GmbH & Co KG, Berlin.
Mittelberg, I. (2019). Peirce's universal categories: on their potential for gesture theory and multimodal analysis. Semiotica 2019: 193–222.
Norris, S. (2013). What is a mode? Smell, olfactory perception, and the notion of mode in multimodal mediated theory. J. Multimodal Commun. 2: 155–170.
Ortega, G., Sümer, B., and Özyürek, A. (2017). Type of iconicity matters in the vocabulary development of signing children. Dev. Psychol. 53: 89–99.
Pérez-Sobrino, P. (2016). Multimodal metaphor and metonymy in. Advertising: a corpus-based account. Metaphor Symbol 31: 73–90.
Pun, B.O. (2008). Metafunctional analyses of sound in film communication. In: Unsworth, L. (Ed.), *Multimodal semiotics: functional analysis in contexts of education*. A&C Black, London, pp. 105–121.
Reijnierse, W.G., Burgers, C., Krennmayr, T., and Steen, G.J. (2018). DMIP: a method for identifying potentially deliberate metaphor in language use. Corpus Pragmat. 2: 129–147.
Rossolatos, G. (2013). Rhetorical transformations in multimodal advertising texts: from general to local degree zero. Hermes-J. Lang. Commun. Bus. 50: 97–118.
Rossolatos, G. (2014). Conducting multimodal rhetorical analysis of TV ads with Atlas. ti 7. Multimodal Commun. 3: 51–84.
See (n.d.). In the Cambridge dictionary, Available at: https://dictionary.cambridge.org/dictionary/english-spanish/see.

See (n.d.). In the Merriam-Webster dictionary, Available at: https://www.merriam-webster.com/dictionary/see.
Šorm, E. and Steen, G.J. (2013). Processing visual metaphor: a study in thinking out loud. Metaphor Soc. World 3: 1–34.
Stampoulidis, G. and Bolognesi, M. (2019). Bringing metaphors back to the streets: a corpus-based study for the identification and interpretation of rhetorical figures in street art. Vis. Commun. 0: 1–35.
Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T., and Pasma, T. (2010). *A method for linguistic metaphor identification: from MIP to MIPVU*. John Benjamins, Amsterdam and Philadelphia.
Steen, G. (2001). A rhetoric of metaphor: conceptual and linguistic metaphor and the psychology of literature. In: Schram, D. and Steen, G.J. (Eds.), *The psychology and sociology of literature: in honor of Elrud Ibsch*, Vol. 35. John Benjamins, Amsterdam and Philadelphia, pp. 145–164.
Steen, G. (2002). Identifying metaphor in language: a cognitive approach. Style 36: 386–407.
Steen, G. (2008). The paradox of metaphor: why we need a three-dimensional model of metaphor. Metaphor Symbol 23: 213–241.
Steen, G. (2009). From linguistic form to conceptual structure in five steps: a procedure for metaphor identification in discourse. In: Brone, G. and Vandaele, J. (Eds.), *Cognitive poetics*. Mouton de Gruyter, Berlin and New York.
Su, C., Huang, S., and Chen, Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. Neurocomputing 219: 300–311.
Van Dijk, J., Kerkhofs, R., Van Rooij, I., and Haselager, P. (2008). Can there be such a thing as embodied embedded cognitive neuroscience? Theor. Psychol. 18: 297–316.
Wicke, P. and Bolognesi, M. (2020). Emoji-based semantic representations for abstract and concrete concepts. Cognit. Process. 21: 615–635.
Wilks, Y. (1978). Making preferences more active. Artif. Intell. 11: 197–223.
Wu, L.L. and Barsalou, L.W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. Acta Psychol 132: 173–189.
Yves Saint Laurent [YSL Beauty] (2015). Black Opium – director's Cut [video file], Available at: https://www.youtube.com/watch?v=a4l2Fuj7L7U (Accessed 18 December 2019).
Zbikowski, L.M. (1997). Conceptual models and cross-domain mapping: new perspectives on theories of music and hierarchy. J. Music Theor. 41: 193–225.