COSMOROE: A cross-media relations framework for modelling multimedia dialectics

Article *in* Multimedia Systems · November 2008

DOI: 10.1007/s00530-008-0142-0 · Source: DBLP

Katerina Pastra
Athena-Research and Innovation Center in Information, Communication and Knowledge Technologies

Multimedia Systems (2008) 14:299–323
DOI 10.1007/s00530-008-0142-0

REGULAR PAPER

COSMOROE: a cross-media relations framework for modelling multimedia dialectics

Katerina Pastra

Received: 27 February 2007 / Accepted: 19 August 2008 / Published online: 13 September 2008

Abstract Though everyday interaction is predominantly multimodal, a purpose-developed framework for describing the semantic interplay between verbal and non-verbal communication is still lacking. This lack not only indicates one's poor understanding of multimodal human behaviour, but also weakens any attempt to model such behaviour computationally. In this article, we present COSMOROE, a corpus-based framework for describing semantic interrelations between images, language and body movements. We argue that in viewing such relations from a message-formation perspective rather than a communicative goal one, one may develop a framework with *descriptive power* and *computational applicability*. We test COSMOROE for compliance to these criteria, by using it for annotating a corpus of TV travel programmes; we present all particulars of the annotation process and conclude with a discussion on the usability and scope of such annotated corpora.

Keywords Cross-media relations · Image–language–movement interaction · Multimedia semantics · Multimedia dialectics · Multimedia discourse

1 Introduction

Everyday situated communication is predominantly multimodal; humans express their intentions using multiple modalities and they understand others' intentionality as expressed in multimodal ways. The ever growing interest in developing multimedia systems and multimodal interfaces for intuitive human–machine interaction is rather expected. Such technology that will understand and reproduce multimodal human behaviour—an embodied conversational agent (ECA), a robot, or an intelligent system in general—requires mechanisms for cross-media decision making. It requires mechanisms that will be able to understand and reproduce the semantic interplay between different media in multimedia discourse.

Automatic indexing, retrieval and categorisation of audiovisual data or data coming from different media sources (e.g., web text, TV programmes, radio files, etc.) need mechanisms that go beyond *single-media content* analysis to *multimedia* and *cross-media* semantics [48,49]. On the other hand, the automatic generation of audiovisual presentations and summaries needs the same mechanisms for media *allocation* and *coordination* [9] decisions,[1] to sustain coherence and cohesion in communication.

In other words, intelligent multimedia systems of a wide range of applications [47,51] require cross-media association mechanisms, that analyse and generate semantic links between different media and modalities. Though many attempts for building such systems have been made, the development of the specific association mechanisms relies heavily on human intervention [47,51].

In fact, though there have been some attempts to describe the different facets of such semantic interplay, one still lacks a "language" to talk about multimedia dialectics. One lacks a descriptive framework for cross-media relations. This was indicated in the 1990s [59] and it still remains largely true,

Communicated by B. Bailey.

K. Pastra (✉)
Department of Language Technology Applications,
Institute for Language and Speech Processing,
Artemidos 6 and Epidavrou, 15125 Athens, Greece
e-mail: kpastra@ilsp.gr

[1] That is for deciding on which modalities should be used to express specific pieces of information more effectively, and for generating cross-media references that signal and smooth the interplay between media.

especially if one considers the characteristics such framework should have the following:

- *Descriptive power*: the framework should be general enough to describe the interaction between any media-pair (e.g., image–language, gesture–language, etc.), irrespective of domain and genre in which this interaction manifests, and of the specific modalities involved (e.g., sketch–text, video footage–speech, etc.), and
- *Computational applicability*: it should be suitable for modelling rather than merely describing media-interaction, it should guide the computational modelling of multimedia dialectics as evidenced in multimodal human behaviour.

These two criteria are rooted within the very same need for intelligent, multimedia systems, since the development of such systems dictates that any attempt to describe cross-media interaction relations should have the following:

- *Wide coverage*, so that it scales beyond specific media pairs.
- *Wide scope*, so that it both deepens one's understanding of multimedia dialectics and facilitates its computational modelling.

In this article, we present COSMOROE, a corpus-based, descriptive framework of image, language and body movement relations that attempts to satisfy the two criteria mentioned above. First, we review related work on image–language and gesture–language dialectics that was undertaken within Artificial Intelligence and Semiotics. Then, we continue on suggesting the COSMOROE relation set which stemmed from an analysis of multimedia human behaviour as captured in multimedia corpora. The set advocates a message-formation rather than communicative goals perspective in describing general and computationally tractable cross-media relations. To illustrate the coverage and computational applicability of the suggested set, we describe our work on employing COSMOROE for annotating a corpus of TV travel programmes with the objective of training and testing cross-media decision algorithms. Last, we conclude with a discussion on the use of and need for similarly annotated multimedia corpora.

2 Perspectives on cross-media semantic relations

The interaction of different media-pairs has been explored from Semiotics, Cognitive Science and Artificial Intelligence perspectives; however, each discipline has a different interest on the issue. Semiotics focuses more on the characteristics of each medium and its use as a semiotic system; semantic interaction among semiotic systems has been studied less thoroughly, mainly in specific genres of discourse, such as advertising, news, literature and education material (cf. for example the related survey in [33]). In Cognitive Science, it is learning, memory and perception aspects of unimodal versus multimodal discourse that are in focus within media-pair studies, rather than the semantic relation(s) between the pieces of information expressed by each medium (consider, for example, Reinwein's[2] online archive of experimental studies on image–language relations). The vast majority of these studies report on cognitive experiments that assess one's understanding when exposed to unimodal or bimodal information and the success of tasks performed by users when given instructions only through language or through pictures or a combination of these two. In Artificial Intelligence (AI), the development of intelligent multimedia systems has led researchers to study such relations for a principled design of the corresponding decision mechanisms [39]; therefore, there is a strong computational modelling interest in their view of the relations, which goes beyond the descriptive character of the Semiotic studies.

In this section, we will review research from AI on the semantic relations of two media pairs, i.e. images and language, and gestures and language, and will correlate it to semiotic perspectives on the issue; the objective is to explore the possible cross-fertilisation between these perspectives for reaching a descriptive framework of such relations that will satisfy the criteria mentioned earlier.

2.1 Image–language interaction relations in AI

To identify multimedia interaction relations for the needs of their intelligent image–language multimedia systems, some developers have undertaken ad hoc studies of non-digitised image–text multimedia documents [38,39]. The studies differ in many respects: they explore different visual modalities (e.g., information graphics, two-dimensional drawings, etc.) and document genres (e.g., economics textbooks, technical instruction manuals, etc.); the level of granularity of the relations they identify varies considerably, and the direction of the relations they define too varies (they are mainly text-centric, but some of them define image-centric relations too). Table 1 gives an overview of the particulars of these studies.

Some of these interaction relation studies have also attempted to indicate general types of information usually expressed by specific modalities. For example, it has been argued that negation is usually expressed through language, while spatial relations/configuration is more precisely expressed visually [1,21]. These findings are only *implicitly* correlated with the interaction relations identified in the studies. Furthermore, the same studies have also looked at cross-modal references [1,2,21,23], without correlating these

[2] http://www.images-words.net (last accessed April 2008), in French.

Table 1 Image–language relations in AI studies

Work	WIP system [1]	Multimedia argumentation [23]	Postgraph [15,16,20]
Modalities	2D drawing-text	2D information graphics-captions	2D information graphics-captions
Corpus	Manuals for espresso machines	Textbook, report/newsletters	Textbooks
Relations (text to image)	Attention invocation	Elaboration	Focusing
	Elaboration	Restatement	Identification/discrimination
	Explanation	Summarisation	Summarisation
	Labelling	Justification	Justification
	Background provision	Evaluation	Evaluation
		Interpretation	Interpretation
Relations (image to text)	Elaboration		
	Clarification		
	Context provision		
	Elucidation		

Table 2 The RST textual discourse relations

Relation categories	Subject–matter	Presentation	Multinuclear
Definition	They inform the hearer	They affect, act upon the hearer	Both segments nuclei (equal status)
Number of subcategories	15	10	7
Example subcategories	Elaboration	Background	Sequence

findings with the interaction relation sets built either. While undertaken for dealing with very specific and diverse discourse genres, applications and modality types, the studies seem to share a—more or less—common glossary for expressing image–language relations. This is no coincidence, since all studies mentioned above adopt a "communicative goal" perspective in the relations they indicate. Some of them have actually adopted the relations identified for textual discourse within the rhetorical structure theory (RST) [32].[3] Other studies use different wording but actually denote similar relations (cf. for example the "restatement" and "elucidation" relations in Table 1 which refer to the same thing, i.e. the repetition of the same information by each modality without adding any extra information). However, are such relations—and RST in particular—appropriate for describing multimedia dialectics?

2.1.1 Using RST for multimedia semantics

Rhetorical structure theory [32] is a framework for describing relations between text segments beyond the clause-level. Its relations are mainly defined in terms of the intended effect of a text segment combination on the reader; furthermore, the definitions include constraints on the "nucleus" (N) text segment and the "satellite" (S) text segment that participate in the discourse relation. The relations are grouped into three categories: *Subject–matter* relations that are used to inform the hearer of something, *Presentation* relations that are used to affect/act upon the hearer and *Multinuclear* relations in which the related segments have equal status (they are both nuclei), i.e. they both contribute to the message equally, one is not subordinate to another (as in the other two relation types). Each of these relations has a number of subrelations (cf. Table 2 with information on the number of sub-relations and corresponding examples).[4] The framework includes both semantic relations (i.e. informational relations pertaining to the content of a text) and relations that express how text affects the reader to accept what is written (e.g., providing background information, justification, evidence, motivation, etc.) [42,43]. The former are mainly intra-sentential, lexically conventionalised and signaled, while the latter are inter-sentential and in many cases not explicitly signaled in discourse [44].

Rhetorical structure theory has been criticised for lack of descriptive power in capturing intentionality [42,43], for

[3] RST defines interaction relations between text segments.

[4] Elaboration: the S provides additional detail to N being, e.g., the member of a set, the part of a whole, the attribute of an object. Background: the S provides background information to N, which is needed for N to be sufficiently comprehended. Sequence: succession relation between situations expressed in the two nuclei.

mixing informational and presentation relations[5] and has been shown difficult to model computationally [11]. However, it has been used extensively in a number of applications in Natural Language Processing (e.g., text generation) and beyond (e.g., discourse analysis) [57]. In some cases, it has also inspired the development of similar frameworks of rhetorical relations *across text documents* (cf. for example [52]). In intelligent multimedia systems, RST has been used as a language to describe image–text relations [1,7], but no attempts for *automatic identification* of such image–text relations or for their systematic use in describing cross-media relations have been reported.[6]

In lacking a language for describing multimedia dialectics, the use of a well-established RST is natural. However, RST has been formulated for describing rhetorical relations in *textual discourse* rather than multimedia discourse and one needs to be aware of a number of characteristics of the framework that question its suitability for modelling multimedia dialectics:

– *The nucleus versus satellite distinction*

The notion of nuclearity has been reported to negatively affect inter-annotator agreement in annotating text corpora with RST relations [11]. Deciding on which text segment is the nucleus (i.e. it carries the most important information), and which is the satellite (i.e. it carries less significant information) is highly subjective and actually is dependent on context and interpretation perspective. Using the notion of nuclearity to define relations between segments introduces fuzziness, which hinders the computational modelling of such relations. However, in shifting this paradigm to multimedia discourse, one encounters problems that go beyond subjectivity or fuzziness:

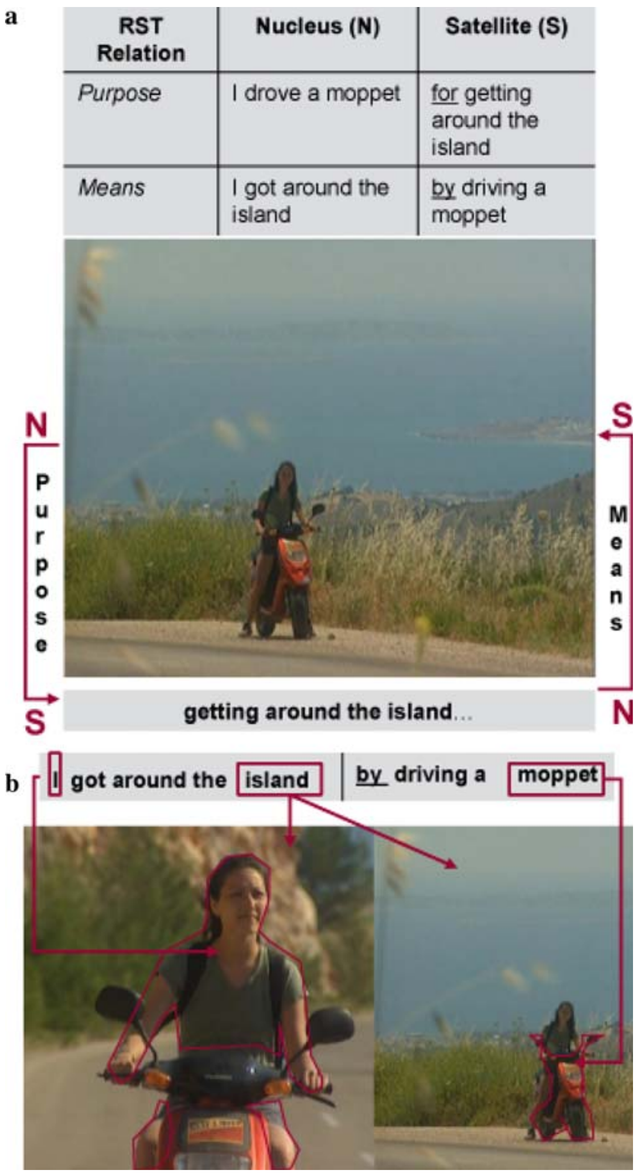
1. *The nucleus versus satellite distinction relies on a single, unique message reading directionality*, i.e. it relies on the fact that the text- or speech-only message manifests itself linearly in space and time; namely, one language segment comes necessarily after another, the nucleus preceding or following the satellite. For example, the RST presentation-relation of *preparation* is defined as one in which the satellite precedes in discourse to make the reader more ready or interested to read the nucleus. However, the multimedia discourse is totally different in the way that it manifests itself. Dynamic modalities in a multimedia document are mainly parallel in time and space (cf., for example, video and audio in a documentary), while static ones are perceived linearly but not necessarily in a strictly predetermined, unique order (cf., for example, illustrated web documents, in which one may focus first on the photograph and then read the text, or the other way around).
2. *Rhetorical structure theory relies—in some cases—on lexical cues and specific syntactic patterns* for distinguishing the nucleus from the satellite and for determining the exact relation that holds between text segments. Consider for example a number of clause subordination cases introduced with prepositions/prepositional phrases that indicate relations such as *purpose* (e.g., "in order to"—the satellite expresses the purpose of the nucleus), the conditional "if", etc. Such cues for distinguishing between nuclei and satellites in multimedia discourse (when the nucleus is one medium and the satellite another) are not present.[7]
3. *The nucleus versus satellite distinction presumes that the segments are comparable in size*, for example, the RST relation of *restatement* is defined as one in which the satellite restates the nucleus and both of them are of "comparable bulk", while that of *summary* requires that the satellite carries the same content as the nucleus but is shorter in bulk [32]. However, the units of the modalities that are engaged in a multimedia discourse relation may belong to different levels of granularity, i.e. the interaction may be between, e.g., a whole image and a word, or an image part/object and a phrase; therefore, the semantic links between modalities may be at any level of discourse, that being a lexical, sentential or higher order one, in contrast to language-only discourse in which discourse relations take place between smaller or larger clusters of clauses. Obviously, any notion of "comparable bulk" is simply non-applicable in cross-media interaction.

Figure 1a, b provide an illustration of the above criticisms. In particular, Fig. 1a presents the case of employing RST in textual versus multimedia discourse: at the top of the figure, one may read two examples of the "Means" and "Purpose" RST relations between text segments. The sentences "I drove a moppet for getting around the island" and "I got around the island by driving a moppet", though largely similar in content can clearly be identified as examples of the two different

[5] Compare [44] for a review on suggestions on the number and type of relations that should be included in a RST, as well as [54] on a suggested criterion for including a relation in such theory.

[6] In a few cases, within multimedia system applications, RST trees that depict the rhetorical structure of a multimedia message have been manually crafted and automatically pruned to satisfy layout or other document generation needs [6,12]; this use of RST remains at a knowledge-representation level. In other cases, RST relations among transcribed speech segments of audiovisual files are manually identified and then used for video generation and retrieval applications [30,53].

[7] Even if, e.g., one medium expresses the purpose of an action denoted by another medium, this type of cues are not there; the relation is implicit in that only pragmatics and the correlation of the media in time or space point to the specific relation.



RST Relation	Nucleus (N)	Satellite (S)
Purpose	I drove a moppet	for getting around the island
Means	I got around the island	by driving a moppet

Fig. 1 RST problems in multimedia discourse

RST relations, mainly due to the prepositions introducing the subordinate clauses. These prepositions (underlined in the figure) also make clear which text segment is the nucleus and which one is the satellite.

Taking such example in multimedia discourse, one may think of a situation in which the notion of "someone driving a moppet" is depicted visually, in a photograph, and the notion of "getting around the island" is expressed through text, as, e.g., a caption of the photograph. While in textual discourse, syntax and lexical cues determine clearly the relation between the two units and their role as nuclei or satellites, in multimedia discourse no such clues are present. The intended relation between the photograph and the caption can only be assumed by the larger context or situation in which this annotated photograph is being used. As it is, it may stand to denote either a "Means" relation or a "Purpose" relation, depending on whether one considers the text or the image to be the nucleus of the message. The directionality in reading this multimedia document is not unique, subtle cues denoting the relation and the nucleus versus satellite distinction are not present.

Figure 1b presents another situation in which the sentence "I got around the island by driving a moppet" may be uttered in a TV travel programme by the presenter while she is actually doing so. In this case, the multimedia discourse consists of the related video/visual scenes and the utterance, both aligned in time, but with the visual scenes extending beyond the end-time stamp of the utterance. In such case, we have a number of "Restatement" relations between text and images; "I" corresponds to the image of the presenter (best viewed in a close-up keyframe, but tracked through the whole visual scene), the "moppet" corresponds to the image of the moppet (best viewed in a specific keyframe, but tracked through the whole visual scene), the "island" corresponding to the background of all frames of the visual scene, and the notion of "driving" corresponding to the whole visual scene, in which an AGENT ("I") drives (motion path from frame to frame/trajectory path of moving agent and object) an OBJECT ("moppet") in a specific PLACE ("island"). This example shows that the RST "restatement" relation holds between units of non-comparable size, units that cannot be distinguished into nuclei and satellites in any way, units that one does not gain anything by even attempting to characterise as nuclei or satellite.

– *No compliance with media characteristics*

Lack of focus indicators and specificity have been indicated as distinguishing characteristics of images [8]; the former, refers to the fact that images have very limited visual means for indicating the salience of the entities they depict. Specificity—on its turn—refers to the fact that images are exhaustive representations of what they stand for, even in cases when their reference object is intended to represent a class of entities rather than a specific individual; in other words, images cannot indicate any type–token distinctions using visual means [25].

In contrast to images, natural language expressions are considered to be able to go beyond things that can be depicted to abstract ideas and reasoning and as Minsky has also indicated, attention to details or focus is controlled in language [41]. Language has a meta-language function that assists in clarifying the level of abstraction of what is expressed. On the other hand, language has no direct access to the physical status of its reference objects as images do; it is in no way tied to sensorimotor representations; it can only indirectly refer to specific instances of physical objects (through, e.g., proper names, deictic references, etc.) [46,47].

Taking these media characteristics into consideration, one realises that many RST relations do not apply to images or language at all; for example, the *elaboration* relation is defined as one in which the satellite presents additional detail about the content of the nucleus, such as the member of a set, an instance of an abstraction, an attribute of an object, something specific in a generalisation. However, images always provide such extra information, by nature, even if this extra information is not important in discourse. Should one consider every image–language relation as one of elaboration?

In Fig. 1b, the nature of the images is such that one gets a number of extra pieces of information not mentioned explicitly in the textual discourse (e.g., the exact characteristics of the moppet, the characteristics of the presenter, whether she wore a helmet or not, the combination of mountainous landscape and the sea in the island, etc.). In that sense, all these pieces of information could be thought of as "elaborations" to what the utterance mentions; however, they could also be thought of as coincidental/non-intentional pieces of information that one unavoidably gets when visualising a situation, pieces of information that should not be linked to the utterance in any way. One needs to be aware of the characteristics of each modality to decide on the correlations to be made, and this must be reflected in the relation framework used. Naturally, this is not the case in RST, since the theory was not developed for describing multimedia dialectics.

2.2 Image–language interaction relations with a semiotics perspective

Semiotic frameworks of image–language interaction relations introduce subjective descriptors for classifying or defining the relations too (cf. Table 3). Barthes [5], for example, identifies three image–language relations, which incorporate a notion of a modality's "contribution" to the message, one reminiscent of the nucleus versus satellite distinction in RST. Following Halliday's logico-semantic relations for clauses and Barthes' notion of "contribution", Martinec and Salway [37] identify another triplet of non-mutually exclusive image–language interaction relations: *equality*, *expansion*[8] and *projection*.

In elaborating on the "equality" relation, the authors mention that modalities may contribute equally to a message, no matter whether one is dependent on another. This claim supports our arguments *against* the use of any "contribution" criterion in a cross-media relation framework. However, in order to explain the "non-mutually exclusive" character of their relations, Salway and Martinec seem to use the criterion, introducing, this way, contradictions in their framework.[9] On top of that, in their attempt to define precisely the relations they have identified, the authors use a "general versus specific" distinction to describe different cases of elaboration. However, how can one decide what is general and what is specific, and which modality is more general than the other? This contradicts the very nature of each medium (cf. Sect. 2.1.1).

Last, in her taxonomy of image–text relations Marsh [33] compiles a set of 49 relations reported within studies in a number of different disciplines (e.g., education, journalism, etc.) all of which have a semiotics perspective. She classifies the relations into ones in which images have a *close relation* to text (e.g., describe, reiterate, concretise), *little relation* to text (e.g., decorate, elicit emotion, etc.) and ones that go *beyond* text (e.g., emphasize, compare, etc.). In this work, it becomes clear how much tied to an "intentionality" (communicative goals) perspective all such attempts are to describe image–language relations are. It is also evident, that criteria such as "closeness in meaning" are almost impossible to define in a way that will allow the computational modelling of such relations.

2.3 Gesture–language interaction relations in AI

While most image–language interaction studies involved multimedia documents, the analysis of gesture–language interaction has focused on real-time use of interacting media by humans. These are captured in corpora of multimodal human–human or human–computer interaction sessions, in which humans make use of gestures and language (speech) to refer to the commonly shared visual space, to things that are not physically present or things depicted on screen. In such cases, the analysis of the images involved (graphics on screen or shared visual space) have been largely taken for granted, which has resulted in a considerably less number of corpora, tools and studies for the image–language pair [51]. It is, therefore, no surprise that a number of systematically built descriptive gesture–language interaction relation sets exist in the literature and actually date back to the late 1990s (cf. Table 4 for a concise view of such relation sets).

[8] Subcategories: *Elaboration* = detailed description, one or both modalities are general, *Extension* = addition of information, *Enhancement* = one modality qualifies another with regard to space, time, cause.

[9] Consider, for example, the case of equal/independent and expansion relations in the same image–language interaction case that the authors claim one may encounter; how can the modalities express meanings that are independent (do not collaborate in forming a larger syntagm) and at the same time one modifies another adding extra information?

Table 3 Image–language relations from a semiotic perspective

Work	Barthes [5]	Martinec and Salway [37]
Relation categories	Anchorage = text supports images	Equality = equal contribution to a message
	Illustration = images support text	Expansion = there is dependency between modalities
	Relay = equal contribution to the message	Projection = meta-information on verbal and mental processes involved

Table 4 Gesture–language relations in AI

Work	Relations
TYCOON [35]	Complementarity, redundancy, transfer, concurrency, specification, equivalence
Multimedia score [31]	Repetition, addition, substitution, contradiction
CoGest [24]	Equivalence, intersection, subset, no-intersection

TYCOON, one of the very few attempts to build a computational interaction descriptive framework, is related to gesture–speech interaction and has been used almost exclusively for the analysis of the interaction of this media pair [34–36]. TYCOON defines six interaction relations that one could group into three clusters each one corresponding to a different perspective: *Complementarity* and *redundancy* are determined according to the contribution of each modality to the formation of the message. In complementarity "different modalities convey different chunks of information...(with some common and some different attributes) that need to be merged", while in redundancy both modalities "express the same information", which also needs to be merged [35]. One would expect further elaboration on how the modalities complement each other, or in what way they express the same information, given their inherently different characteristics, but the framework does not actually focus on these aspects.

The strong *computational analysis* perspective of the relations identified is more evident in the case of the other two relations of the framework, i.e. *transfer* and *concurrency*, which capture different processing cases of the modalities (sequential and parallel, respectively). On the other hand, *specification* and *equivalence* refer to *media effectiveness*-related information, since they indicate whether specific types of information are expressed consistently with only one modality or not. Therefore, a specific multimodal discourse segment may actually denote three of these relations at the same time, corresponding to the three different perspectives/criteria used to determine modality co-operation. For example, the multimodal segment, "show me the hospital"—"CLICK" (pointing gesture on the icon of one of the hospitals depicted on screen), is an example of complementarity/merging between language and gestures. However, there is also a concurrency relation, if the gesture takes place at the same time when the phrase "the hospital" is uttered and an equivalence relation, if information regarding hospitals is found to be sometimes referred to through speech and some other times through gestures.

In another work, the Multimodal Score approach for analysing multimodal human behaviour [31], four semantic functions between gestures/body movements and corresponding speech are indicated: *repetition, addition, substitution, contradiction*. The relation set seems influenced from the notion of "mutation operators" in science and is quite simple and straight-forward. However, it describes the processes rather than the semantics of interaction between modalities. As such, its relations are considered mutually-exclusive, when actually the cases they correspond to may not be exclusive. For example, one may utter the phrase: "I am going back home" while using—at the same time—an iconic hand-gesture of "walking". In one sense, the textual "going" is repeated through the gestural "walking", but the gesture adds also more information, specifying the means used for "going". Furthermore, the substitution relation refers to "cases when a gesture replaces a word not uttered at all", but actually even the above mentioned example may be considered a case in which the gesture replaces the non-uttered phrase "on foot".

In CoGest, a gesture–language annotation tool [24] for conversational gesture-generation, four types of meaning interaction were indicated: *equivalence, intersection, subset* (modest contribution), *no intersection* of meanings between the interacting modalities. The relations rely on set-theory concepts and refer to the size of contribution and degree of meaning overlap between the modalities; however—as the developers admit themselves—such criterion is fuzzy and highly subjective.

In contrast to image–language studies, gesture–language studies do not attempt to express "intentionality" through the interaction relations they identify. They are in search of a different perspective for describing such relations, and they seem closer to terms that describe the interaction processes for meaning formation.

2.4 Gesture–language interaction in semiotics

Turning to theoretical studies of gesture–language interaction in human–human communication, theories and examples of how gestures and speech form part of the same, coherent conversation plan have recently shed light to the interaction between the two media. Though some suggestions on the cognitive mechanisms involved in this process have been made [40], we will briefly look into the descriptions of the functional role of gestures within such interaction and their contribution to propositional content. This role has been explored by researchers since many decades and a number of different—more or less commonly agreed—types of gestures have been defined.[10] Recently published work by Adam Kendon [26], focuses even more on the role of gestures in discourse and its interaction with language in particular; gestures are found to function in four distinct ways for the following:

- *Pointing*: used for deixis
- *Content representation*: they enact the objects/events mentioned in the utterance or their attributes
- *Interaction regulation*: e.g., define speaker turns
- *Meta-discourse signaling*: e.g., mark aspects of discourse structure, display speech act, etc.

Their contribution to propositional content is further explained and illustrated through examples in which gestures *specify* what is being said (e.g., express manner), *add related meaning* (i.e. what emblematic gestures do, they actually have equivalent meaning), *exhibit* what is being said (e.g., enact the object spoken of), *display* object properties or spatial information and serve as the *objects of deictic references*.

Though different in nature, images and gestures seem to interact in a similar way with language. Images realise the object of deictic reference expressed in language, while gestures are the carriers of the reference itself (i.e. images provide the content of the deictic, gestures enact the deictic reference). Still, one of their common functions is that they complement language adding similar dimensions to the message communicated (e.g., specify manner, spatial information, etc.) or they provide further means of expressing the same thing (in case of meaning equivalence). In both cases, gestures are used to represent content, as images do by nature. We would say that gestures actually create a *mental image* of reality, to compensate for the absence of the latter, i.e. they usually function in this way with speech when one describes something that is not visible in the shared communication space.

In the next section, we will present a framework for describing the semantic relations in which all three media (and their corresponding modalities) engage.

3 The COSMOROE relation framework

The CrOSs-Media inteRactiOn rElations (COSMOROE) framework describes multimedia dialectics based on the analysis of multimedia documents and in particular, the analysis of the interaction between images, language and body movements. It is stripped from any criteria related to the *contribution* of each modality to the multimedia message. It actually looks at cross-media relations from a multimedia discourse perspective, i.e. from the perspective of the dialectics between different pieces of information for *forming a coherent message*. Therefore, the relations attempt to capture the semantic aspects of the message formation process itself and thus, facilitate inference mechanisms in their task of revealing (or expressing) intentions in message-formation.

COSMOROE uses an *analogy to language discourse* analysis for "talking" about multimedia dialectics. It actually borrows terms that are widely used in language analysis for describing a number of phenomena (e.g., metonymy, adjuncts, etc.) and adopts a message-formation perspective, which is reminiscent of *structuralistic* approaches in language description. While doing so, inherent characteristics of the different modalities (e.g., exhaustive specificity of images) are taken into consideration.

COSMOROE is the result of a *thorough, inter-disciplinary review* of image–language and gesture–language interaction relations and characteristics as described across a number of disciplines from computational and semiotic perspectives (cf. preceding sections). It is also the result of *observation and analysis* of three different types of corpora for different tasks: the first one was a non-digitised collection of 500 caricatures that covered the work of three different artists with carefully selected, radically different stylistic variations. The images had the form of two-dimensional drawings and the size of the accompanying text varied from object labels and titles to captions and short dialogues/monologues. The analysis of this corpus was done from a semiotics perspective, in an attempt to analyse the language of caricatures, i.e. the image–language interaction in this genre of multimedia documents [45].

The second corpus was a collection of 500 crime-scene photographs and accompanying captions (crime-scene photo albums) and 65 such photographs with spoken (transcribed) captions. The latter resulted from an experiment in a mock crime scene, where crime-scene officers used digital technology to create the multimodal documentation required when visiting a scene [50]. The objective of this analysis was the development of a text-based image-indexing algorithm, and

[10] Emblematic, deictic, iconic, metaphoric and beats—cf. the overview study in [13].

therefore, computational applicability was of primary importance.

The third corpus was a 20-h collection of travel documentaries (half of them videos in English, the other half in Greek), collected in the framework of the REVEAL THIS project [48], and the objective of the analysis was the development of cross-media decision mechanisms for indexing and retrieval of multimedia documents.

There is a variety of modalities covered across these genres (e.g., sketches, photographs, video footage, text, speech, gestures), a variety of language registers (slang, colloquial language, crime investigation terminology), a variety of domains, constraints and needs in forming multimedia messages (e.g., brevity in caricatures, sticking to the facts in crime-scene photographs). However, while originally formulated through analysis of cross-media interaction relations in newspaper caricatures, crime-scene photographs and travel documentaries, COSMOROE was actually elaborated and finalised when shifting to the *TV travel programmes* genre. This is no coincidence, since this genre captures natural, everyday multimodal human interaction and at the same time it is also the product of multimedia document creation (e.g., it contains graphics). This means that it is *rich in media and modalities* and therefore ideal for cross-media relation analysis.

In particular, TV travel programmes usually involve one or more presenters visiting different places, being in contact with the locals, interviewing people, explaining the habits, traditions and way of life in specific locations. Language is used to refer to a variety of things, ranging from tangible things directly depicted in the programme to more abstract concepts. It covers a wide range of concrete and abstract concepts, as is the case in everyday interaction. There is a mix of specific terms and everyday colloquial language that is being used, and there are no strict restrictions in terms of the vocabulary to be used, the length of the descriptions or the visual modalities. Therefore, these audiovisual files include a variety of language modalities (speech and text in the form of subtitles, text in graphics, etc.), image modalities (dynamic images/image sequences, graphics such as maps), gestures and other body movements. In many cases, the files contain section-titles, i.e. captioned frames, in which one may observe modality interaction between an image and its caption, as one would with a static photograph and its accompanying caption. Thus, one is able to observe and annotate a wide variety of *modalities interacting one with another in most possible combinations*.

This genre of audiovisual documents is richer in cross-media relations not only from caricatures and photo-albums, but also from traditional travel documentaries and other TV programmes such as news and sports. In classical (travel or other) documentaries the narrator is usually not present in the video and the whole narration is impersonal. In news, it is mainly the anchorman and the interaction with guest-interviewees that takes place, which means that language and limited gestures prevail in the discourse; images are restricted to depicting the people involved and the venue. It is only in the "reportage" sections of the news that video footage is rich, since it provides information directly related to the content of the news story. Weather and sports programmes also have a very restricted and well-structured format, in which modality interaction is restricted, when compared to travel programmes. For example, in a sports programme, the verbal description of what goes on in the game refers to naming of players and a limited number of events, ones that are well defined and regulated.

Therefore, we have chosen to apply COSMOROE in a corpus of TV travel programmes, for testing its coverage and descriptive power. It is expected that if adequate for this genre, which is rich in cross-media relations, COSMOROE will be applicable to any other type of multimedia documents. In different genres, different COSMOROE relations will be more or less frequent, some of them may be present, others may not, exactly as some modalities may be present in the document and others not. The idea is for COSMOROE to cover all possible cross-media interaction relations, so that one may instantiate any of them according to the needs and characteristics of a particular multimedia document genre. For example, movies/films are comparable to TV travel programmes in the richness of interacting modalities. However, due to their artistic nature they may become at times surrealistic and abstract away from everyday interaction. We expect that in such cases, particular COSMOROE relations will be more frequent than in realistic films (cf. for example the "figurative equivalence" relations explained in the next section).

In what follows, we present COSMOROE through examples from a corpus of TV travel programmes annotated with such relations, provide details on the annotation process that aims at employing this framework for computational purposes (the development of cross-media decision algorithms) and conclude with a discussion on the latter.

3.1 Multimedia dialectics from the COSMOROE perspective

Figure 2 presents the three core COSMOROE relations with their sub-relations. In particular, these relations are as follows:

- *Equivalence* (Multimedia message = X = Y): the information expressed by the different media is semantically equivalent, it refers to the same entity (object, state, event or property). This is a case of grounding language in perception and perception in conceptualisation/language [46]. Drawing an analogy to language discourse,

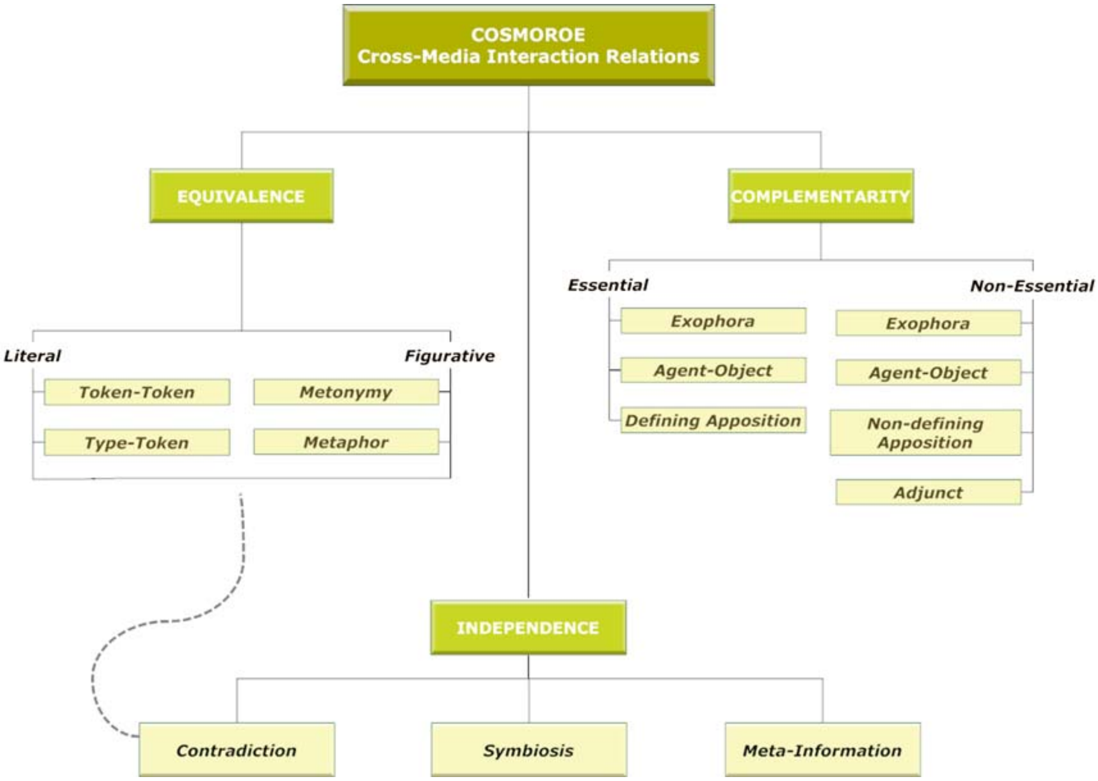


Fig. 2 The COSMOROE cross-media relations

equivalence relations could be thought of as *paradigmatic* relations between modalities.

- *Complementarity* (Multimedia message = X + [Y]):[11] the information expressed in one medium is (an essential or not) complement of the information expressed in another medium. Association signals (e.g., textual indexicals pointing to an image or image part) indicate cases of essential complementarity, while non-essential complementarity is characterised by one medium modifying or playing the role of an adjunct for the other (e.g., an image showing—among others—the means used by the speaker to reach the place she mentions at the corresponding audio stream of the document). Complementarity relations have a *syntagmatic* (syntactic) nature.
- *Independence* (Multimedia message = X + Y): each medium carries an independent message, which is, however, coherent (or strikingly incoherent) with the document topic. Their combination creates the multimedia message. Each of them can stand on its own (it is comprehensible on its own), but their combination creates a larger multimedia message (it is like a conjunction of sentences).

Going further into the subtypes of these core relations, we distinguish four subtypes of semantic equivalence, two of them pertaining to literal equivalence and the other two to figurative. Starting from literal equivalence, two cases in which the media refer to the same entity/action/property have been distinguished: the *token–token* and the *type–token* ones. The distinction is intended to deal with cases in which the media refer exactly to the same entity, uniquely identified as such, and ones in which one medium provides the class of the entity expressed by the other. For example, a person name and the corresponding image of that person, stand in a *token–token* relation, i.e. there is an exact match between what is being said and what is being depicted. Linguistic deictics and the corresponding pointing gestures also stand in a token–token relation, cf. for example the word "there" and an accompanying pointing gesture: they both denote place–direction and actually carry no further meaning by themselves. The token–token relations could be thought of as instructions to an algorithm to look for an almost exact match, when associating the specific media.

On the contrary, Fig. 3a shows a case in which the word "housing" is depicted in a series of frames that present various blocks of flats, other buildings, terrace houses, etc. It is a case of a *type–token* situation, in which one modality (text in the example) refers to a class of entities and the other (sequence of video frames in the example) refers to one or more (repre-

[11] The bracket squares denote that information in Y may sometimes be necessary. This "necessity" indication is not defining for the relation and its subtypes, it just denotes the *strength* of the relation between the modalities.

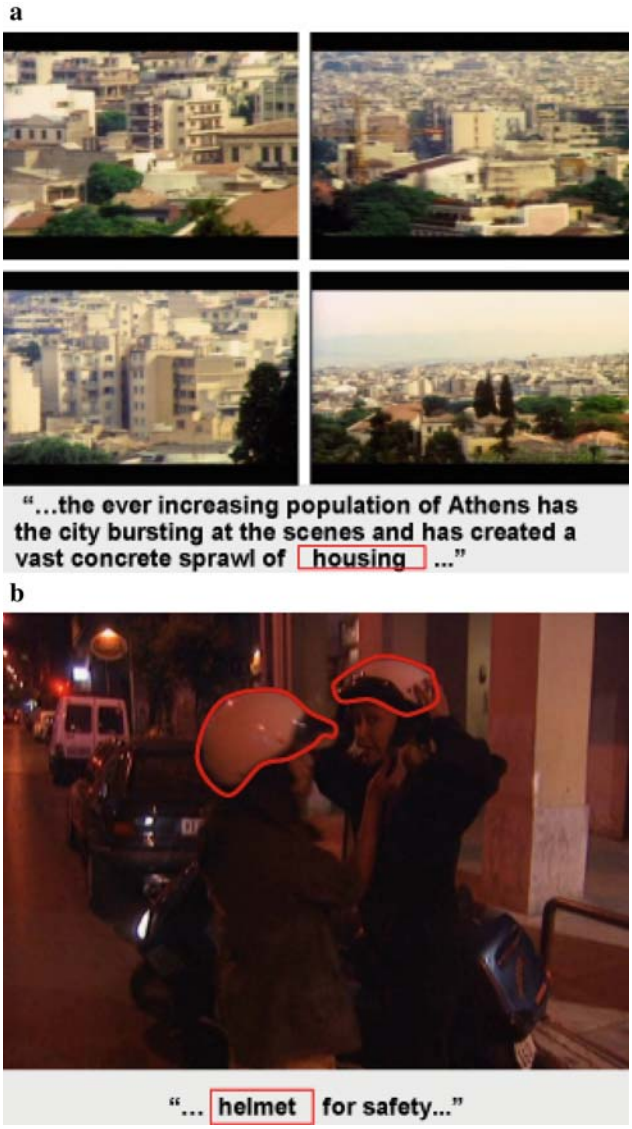

a
"…the ever increasing population of Athens has the city bursting at the scenes and has created a vast concrete sprawl of housing …"
b
"… helmet for safety…"

Fig. 3 A "type–token" equivalence example

sentative) members of the class. Compare also cases of, e.g., the word "furniture" and the image of a "chair", the word "coloured" and the image of something "red"; there is an *IsA* relation between the two referents.[12] The entities linked with such *IsA* relations may be at different levels of abstraction; namely, thinking in terms of a taxonomy, the token may be the immediate hyponym (child) of a concept or may be the hyponym of a number of concepts which are themselves hyponyms of the "type" concept. For example, someone referring to "artefacts" while showing chairs and

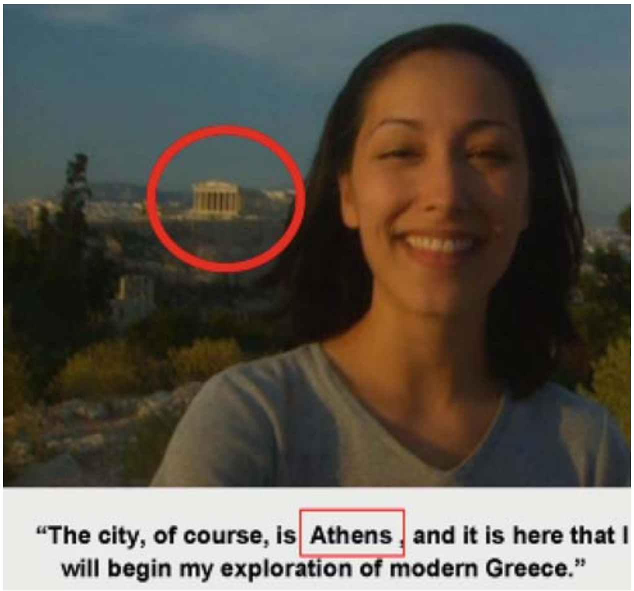

"The city, of course, is Athens, and it is here that I will begin my exploration of modern Greece."

Fig. 4 A "metonymy" example

sofas, draws a type–token relation between a more abstract concept and instantiations of the hyponym concept "furniture", going directly to the hyponyms of the latter. Similarly, referring to, e.g., "helmets" while showing someone wearing a helmet, is a type–token relation, in which a concept (class of entities) is instantiated with a specific type of helmets (it could be any other type of helmets depicted). Figure 3b illustrates such case.

In the other two cases of *equivalence*, i.e. *metonymy* and *metaphor*, we have a figurative association between two different referents, i.e. each modality refers to a different entity, but the intention of the user of the modalities is to consider these two entities as semantically equal. As in language, these two cases are quite different in multimedia discourse too; in metonymy, the two referents come from the same domain, they have the same array of associations and, there is no transfer of qualities from one referent to another.[13] Consider, for instance, the example in Fig. 4, in which the speaker says that she is in Athens; and the background scene depicts the Acropolis (she is close to the Acropolis site). The view of the Acropolis is considered to be equivalent to the view of Athens, Acropolis is a symbol of the city, and this metonymic relation is also evident from the use of the phrase "of course" on the part of the speaker, who considers the identification of this semantic equivalence between what is shown and what is being said evident for any viewer.

[12] It is not only the frame sequence, which is time-aligned with the utterance of the word "housing" that is engaged in a semantic equivalence relation with the word, but also frame sequences (shots) that precede or follow in time; furthermore, the associated groups of frames are not necessarily one after another; they may be "interrupted" with other frames that do not correspond to the specific utterance.

[13] In some cases, one referent is an aspect of the other, or a part of the other, one is a species the other is the genus, one is a material the other is a thing made of this material. In linguistics, such cases are sometimes considered to be cases of Synecdoche; however, we will not differentiate these as a different phenomenon, we will consider all these to be cases of metonymy.

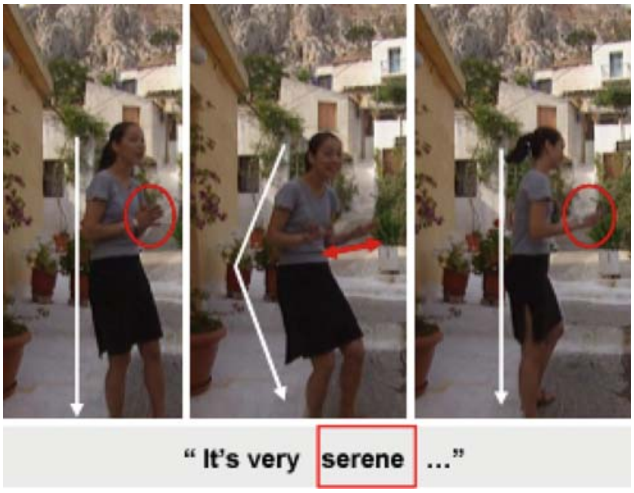


Fig. 5 A "metaphor" example

On the other hand, in cases of *Metaphor*, one draws a similarity between two referents, which actually belong to different domains; there is a transfer of qualities from one to another. Figure 5 illustrates a case of metaphor in multimedia discourse; in this example, the word "serene" is semantically equivalent to a body movement that is sometimes used to denote that something is calm, serene: the body movement consists of a hand gesture and instantaneous bending of the posture (hands touching palms in front of the chest—hands gradually apart on the same level—while the hands are apart the knees bend, lowering the body a bit and then up again with hands back together or hands down).

Going further on to cases of *complementarity*, one distinguishes four different sub-relations clustered into two groups: those in which complementarity between the pieces of information expressed by each medium is *essential* for forming a coherent multimedia message, and those in which complementarity is *non-essential*. In the case of essential complementarity, the meaning of what is communicated is clearly comprehended only when information by all participating media is combined. Explicit or implicit cues *must* be present in discourse for one to characterise indeed the complementary information as essential. Simply put, cases of essential complementarity are all those that "force" one to look for extra information when exposed to the message carried by only one medium.[14] In particular:

– *Essential exophora*

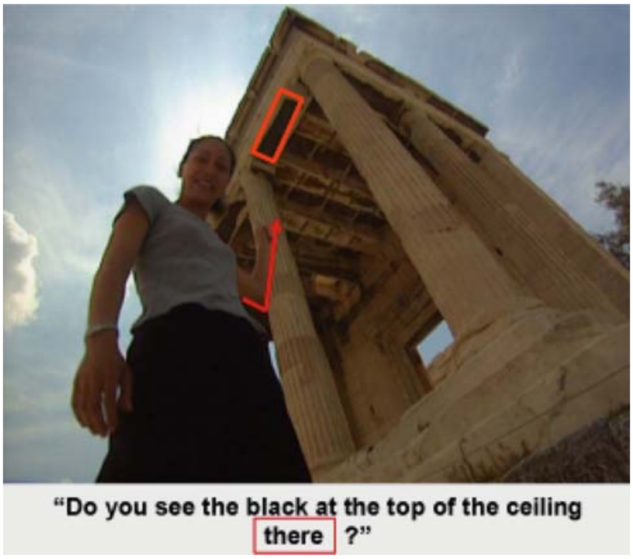


Fig. 6 An "essential exophora" example

Cases of "anaphora" are one in which one medium resolves the reference made by another. Signals of such resolution are present in discourse, e.g., linguistic indexicals or pictorial signs such as arrows pointing to part of an image, or even pointing gestures. The signals indicate a relation between the medium in which they are being expressed (e.g., language) and another medium (e.g., image) that provides the resolution of their reference. For example, the word "this" may point to something pictorial, but its function is just to point to something. It does not express what the thing pointed to is. The latter is information that is provided only by the image; consider, for example, Fig. 6 in which we have two signals: the deictic gesture and the word "there". Both of them signal that somewhere in the context (image region highlighted in red) one will find their reference. The signals themselves are semantically "empty" . The most frequent signals are textual indexicals and deictics, and deictic gestures, which point to an image or image part.

– *Essential agent–object*

In this relation, one medium reveals the subject/agent or object of an action/event/state expressed by another. For example, one may think of a case when someone says, e.g., "they have..." and completes the utterance with a gesture for money; there is an ellipsis phenomenon in the utterance, which signals that somewhere in the context (gesture in this case) one will find the missing argument (i.e. the object of the verb). Figure 7 shows another example of an essential object relation: in a part of the "behind the scenes section" the presenter appears saying to the coachman to "hold"; the utterance is of course highly elliptic and forces one to look at the image, which reveals the object of the verb, i.e. the microphone that is given to the coachman to hold, so that the presenter drives the coach. In this relation, phenomena of ellipsis

[14] Compare, for example cases when one has to look at the TV screen to get information that will complement what one has just heard being said.

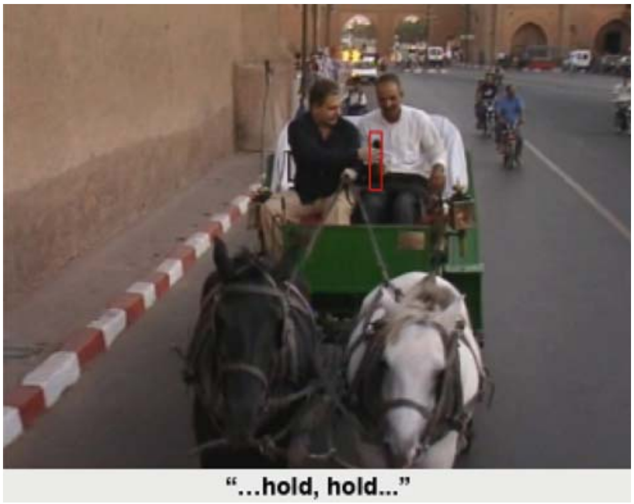


Fig. 7 An "essential agent–object" example

in language point to the fact that complementary information is essential for comprehending the message. While cases of missing objects in the textual part of multimedia discourse are more straight-forward, one may be surprised with the case of missing subjects. It is indeed true that hardly in well-formed speech/text in most languages is the subject of a predicate totally missing.[15] Information on the subject may be evasive (cf. for example passive voice impersonal constructions) but still present. However, consider cases of verbal nominals, use of gerunds, and use of participles in image-captions. In such cases, language is used to focus on the event rather than on the agent, letting the image to fill in this vital piece of information.

– *Defining apposition*

In defining apposition, one medium provides extra information to another, information that identifies or describes something or someone. These are cases of going from something general to something concrete so that an entity is uniquely defined/identified. Consider, for example, Fig. 8 in which "the President of Greece", Kostis Stephanopoulos is uniquely identified through the image of the president at the time the video was filmed. What makes this situation different from the "token–token or type–token equivalence" relation is that it is tied to the specific context and should not be considered as generally valid (i.e. the "President of Greece" is a title that is used for many different people at different time-periods, namely, Mr. Stephanopoulos was the president for a time period but not any more). It is not the same case as in, e.g., the association of one's name with one's photograph, or the association of the word "furniture" with the photograph of a chair, which though crossing over different conceptual levels is not tied to a specific context of discourse.

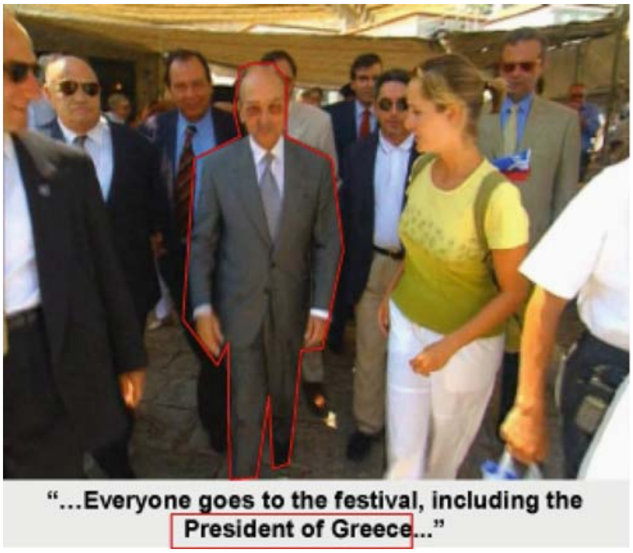

"...Everyone goes to the festival, including the President of Greece..."

Fig. 8 A "defining apposition" example

In the case of *non-essential complementarity*, one medium provides extra information to what the other expresses, information that is not vital for the comprehension of the latter. We distinguish three subtypes of non-essential complementarity:

– *Non-essential exophora*

Cases of anaphora are one in which the entity referred to in one medium is revealed in another, though this is not vital for understanding the message. Figure 9 shows an example of an exophora case, in which the narrator says that "the city is a jumble of the ancient and the modern" and the referent of the nominalised adjectives "the ancient" and "the modern" is revealed in the video footage, showing images of ancient and modern BUILDINGS. The images show the actual (ancient and modern) entities that are being referenced in the utterance; however, no cues are present to show that looking at the image is necessary for understanding the utterance. The general, evasive reference of the nominalised adjectives does not complicate or hinder communication.

– *Non-essential agent–object*

This is a case of one medium providing information on the missing agent or object of an action/state/event expressed by another medium, though this is not vital for understanding the message. In such cases, the missing agent or object is known from the wider communication context or from the shortly preceding multimedia discourse or they are intentionally left vague. For example, consider a case in which

[15] It will be lexicalised (preceding the predicate) and/or expressed through the inflection suffixes of the predicate itself (cf. case of pro-drop languages) or will be agglutinated to the predicate, etc.

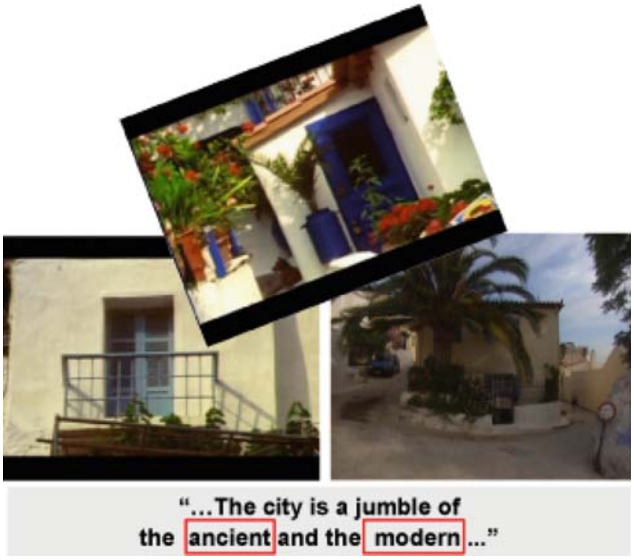


Fig. 9 A “non-essential exophora” example

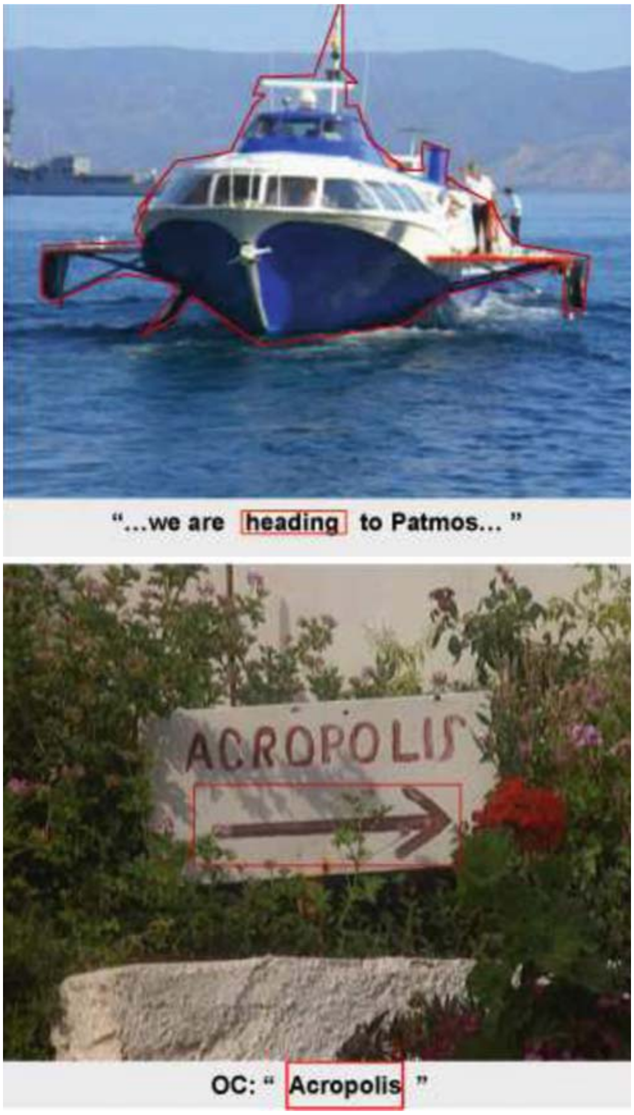


Fig. 10 Examples of “adjunct” cases

a travel documentary presenter says “we spent the whole day shopping”, while the video shows images of clothing and souvenirs, examples of the things they went shopping for. In this case, the complementary visual information provides extra information which is, however, not necessary for understanding the message and is generally implied by the previous discourse (references to shops with famous brands for clothing in a specific area).[16]

– *Adjunct*

This relation denotes an adverbial-type modification (place–position, place–direction, manner). One (or more) media function as adjuncts to the information carried by another medium. Figure 10 presents two examples of such relation in multimedia discourse. In the first one, the presenter of the travel documentary mentions that she is “heading to” an island, while the corresponding images show a flying dolphin (high speed ferry boat); the image reveals the manner used to visit the island; it actually complements the predicate “to head to a place”. In the other example, the arrow depicted in the image complements the inscription “Acropolis” revealing the direction one should follow to reach the Acropolis.

– *Non-defining apposition*

One medium reveals a generic property/characteristic of the very concrete entity mentioned by another medium. Compare for example the case of someone saying that “Mr. Smith was present at the crime scene”, while the corresponding image shows Mr. X, and in particular it shows him wearing a cleaner’s uniform (i.e. Mr. X was a cleaner). In this case, the image reveals information on the occupation of Mr. X that is not mentioned through speech, because it is not related to the main message carried by this medium. One needs to note of course that, by nature, images give much more descriptive information for real world entities than what is mentioned through speech/text (the latter focuses on what is important in discourse, can be elusive and not give details on the appearance of objects, while images are always very specific, they always visualise the shape of something or the colour/hue, etc., and the more realistic/detailed they are the more appealing they also are). In non-defining apposition, we focus on cases in which the extra information provided by the visual

[16] Compare also cases in which one tags his/her photographs bearing in mind that the people who will see them know him/her, so there is no need to repeat every time who is depicted in the photograph, and just concentrate on the action depicted.

modality classifies an entity (e.g., occupation of a person, ethnic origin, etc.).

Last, the relation of *independence* consists of three subtypes:

– *Contradiction*

Usually, in artistic genres of multimedia discourse (e.g., films, newspaper caricatures, etc.). one may find cases of contradiction. Contradiction is the opposite of semantic equivalence, i.e. when one medium refers to the exact opposite of another or to something semantically incompatible; compare for example an image caption saying “our furniture”, while the accompanying image depicts “rocks” (i.e. one has no actual furniture but sleeps/sits, etc., on the rocks). This contradiction relation has exactly the same subtypes as the equivalence relation. The furniture example is a type–token contradiction. A token–token contradiction would be one between the word “Eiffel Tower” and the image of the Parthenon. A metonymy contradiction would be one, e.g., in “I am in Athens”, while the Tower of Pisa appears behind the speaker. A metaphoric contradiction would be one in, e.g., “the giant is here” and the image of a dwarf. In the contradiction relation, the multimedia message may be characterised by irony or humour, which is revealed only through the combination of the pieces of information carried by each medium. In some cases, contradiction may also emerge from editorial mistakes in creating a multimedia document, such as mistakes of video footage choice or footage length for a specific narration segment, or human mistakes in describing entities/situations. An example of a contradiction due to an editorial mistake is shown in Fig. 11, in which the speaker describes an authentic, Moroccan local market in which no tourists go, while the viewers are shown images of the tourist-market of the area (images of tourists in a characteristic souvenir shop).

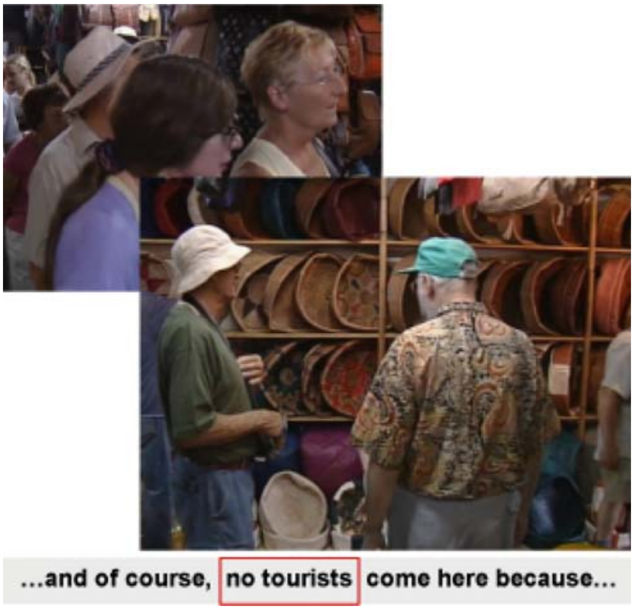

...and of course, no tourists come here because...

Fig. 11 A “contradiction type–token” example

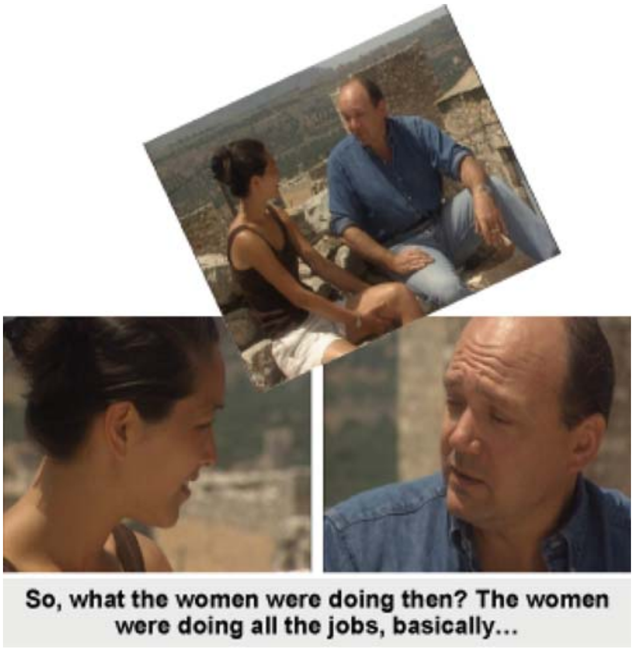

So, what the women were doing then? The women were doing all the jobs, basically...

Fig. 12 A “symbiosis” example

– *Symbiosis*

Symbiosis is a case of different pieces of information being expressed by the media, the conjunction of which (conjunction in time or space) serves “phatic” communication purposes, i.e. one medium provides some information and the other shows something that is thematically related, but does not refer or complement that information in any way. It is just being there for creating a multimedia message. Figure 12 presents such an example, in which the presenter of the travel programme discusses with a historian about the role of women during a specific time period in the villages of a specific place in Greece. What one sees is the speakers themselves. At times, a glimpse to the outdoor area the speakers are located at is shown in the background; the area is generally part of the place for which they give historical information about, but does not show the villages they refer to. In symbiosis, the images that one watch are visual *fillers* that accompany linguistic references to things that are not depicted; their presence is circumstantial (e.g., the result of filming a conversation in a specific setting).

– *Meta-information*

In this case, one medium reveals extra information through its specific means of realisation (or through specific non-propositional types of it); it forms part of the multimedia message, and due to its nature (non-propositional) it stands independently but inherently related to the information expressed by the other media. For example, consider part of a TV travel documentary in which the narrator mentions, e.g., that she is "travelling through the steep mountains" while one watches images from the route through the mountains and the filming is done from within a moving vehicle. This is a multimedia message with verbal information, visual information and visual meta-information. The visual meta-information (i.e. the filming particulars) qualifies the corresponding images (images of the landscape) but also relates to the verbal information by supporting/enhancing the notion of "travelling", (since the filming/the camera is travelling—static camera but on the move); this relation between the verbal information and the filming information is what we call a *meta-information* one. Gestures of non-propositional content may also participate in the multimedia message providing information on how one should parse the message, or on how the interaction between interlocutors is regulated (e.g., a speaker's gesture to prevent the interlocutor from interrupting). In such cases, gestures form also part of the multimedia message; they carry extra information independent from pieces of information expressed by the other media, but nevertheless, inherently related to them. On the part of language, prosody and punctuation in speech and text respectively participate in such meta-information relations (they qualify/modify the language content and may also interact with other media).

3.2 Employing COSMOROE in corpus annotation

COSMOROE was employed for annotating the semantic interplay between different media in a corpus of TV travel programmes. The annotation endeavour has reached 3 h of travel programmes in Greek and one and a half hour of travel programmes in English.[17] The annotators were two postgraduate students in computational linguistics with a linguistic background but no particular knowledge of semiotics, multimedia processing issues and technologies. Their introduction to the annotation task consisted of a brief introduction to media characteristics and multimedia processing technologies as a two-part seminar and some preliminary annotation guidelines and examples of the COSMOROE relations. Their main training took place while they performed the task, i.e. during the annotation of one programme each; questions and clarifications were given on the spot. When they finished their first files (Greek travel programmes), they submitted them to an expert annotator[18] for validation. The result of the validation was a second round of guidelines with more examples, special cases and advice on common mistakes made and how one could avoid them. These enriched guidelines were followed for annotating the second pair of Greek travel programmes, which have also gone through validation, a process which led to an even more refined version of the guidelines and the annotation scheme itself. The English travel programmes have been annotated by the expert annotator directly.

In this section, we will look into the annotation scheme used, the annotation process and environment and we will present details regarding the validation process, the results of inter-annotator agreement and the challenges of the task for the annotators.

3.2.1 Annotation scheme

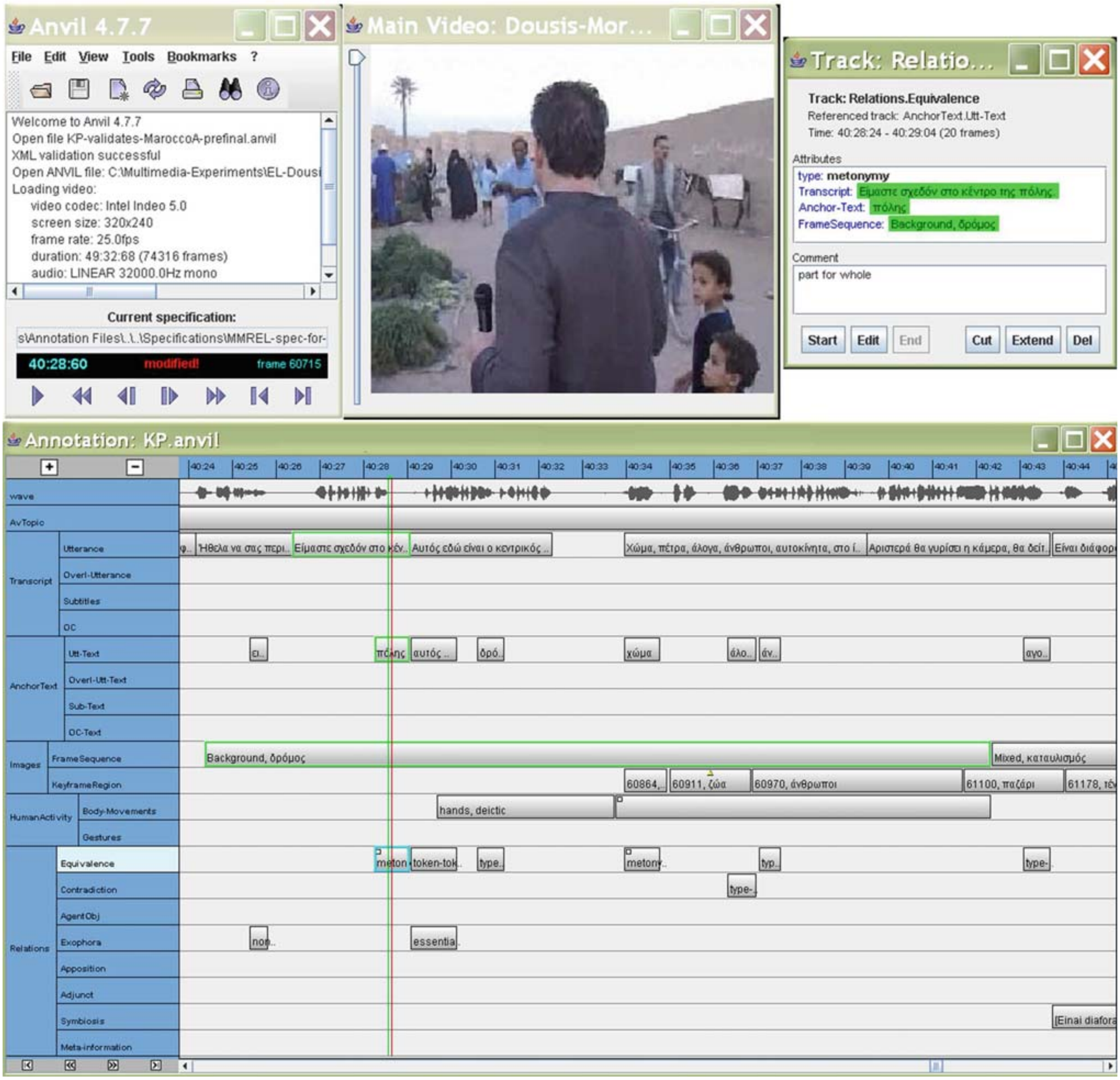
In annotating a corpus with the COSMOROE relations, one needs to employ a multifaceted annotation scheme in which the different media and modalities are being indicated first and then the ones that participate in a relation are singled out and used to populate the relation-annotation template. Therefore, COSMOROE relations link two or more annotation facets, i.e. the modalities of two or more different media. We have developed an annotation scheme in which entities from the speech transcription level (e.g., words or phrases from the speech transcript) and other text-transcription levels (e.g., subtitles and graphical text) are identified as such (their time-offsets are annotated), entities from the body-movement annotation level (gestures and other body movements) are identified as such (their time-offsets are annotated) and entities of the image annotation level (shots and keyframe-regions) are identified as such (their time-offsets are annotated); entities from all these levels participate in COSMOROE relations.

The COSMOROE annotation task is initiated with the manual transcription of the speech stream using the TRANSCRIBER tool [4]. The segmentation of the speech stream into utterances relied on acoustic and syntactico-semantic criteria, i.e. transcription was done at the clause-level, without separating main clauses from subordinate ones or conjuncted/disjuncted clauses unless a distinctive pause did so on the part of the speaker. So, the main criterion was to reflect the uninterrupted speech flow of the speaker in transcription and to avoid *forcing* the clause-level transcription. Distinctive audio events and background music time-offsets have also been transcribed with the tool.

The second step of the annotation process involves the use of the Anvil annotation environment [27], in which

[17] Actually, it is four 1-h programmes in Greek and two 1-h programmes in English; however, we report here the actual duration taking out advertisements.

[18] This was the developer of the COSMOROE framework, and in that sense she is characterised as an "expert annotator" for the task.



Welcome to Anvil 4.7.7
Open file KP-validates-MaroccoA-prefinal.anvil
XML validation successful
Open ANVIL file: C:\Multimedia-Experiments\EL-Dousi
Loading video:
video codec: Intel Indeo 5.0
screen size: 320x240
frame rate: 25.0fps
duration: 49:32:68 (74316 frames)
audio: LINEAR 32000.0Hz mono

Current specification:
s\Annotation Files\..\..\Specifications\MMREL-spec-for-

40:28:60 modified! frame 60715

Track: Relatio...

Track: Relations.Equivalence
Referenced track: AnchorText.Utt-Text
Time: 40:28:24 - 40:29:04 (20 frames)

Attributes
type: **metonymy**
Transcript: Είμαστε σχεδόν στο κέντρο της πόλης.
Anchor-Text: πόλης
FrameSequence: Background, δρόμος

Comment
part for whole

Start Edit End Cut Extend Del

Fig. 13 A screenshot from a COSMOROE annotation session using ANVIL [27,28]

the manual transcriptions are loaded, along with the COSMOROE annotation scheme, for the main annotation task to take place. This annotation environment has been chosen for the task, mainly due to its flexibility and ease of use in defining multifaceted annotation schemes. Figure 13 is a screenshot of an annotation session; when loading the COSMOROE specification file in ANVIL, the annotators see a number of annotation tracks, each one corresponding to different aspects of the video that are being annotated (different annotation levels). In particular, the COSMOROE annotation scheme consists of the following annotation tracks, which appear as soon as one loads the scheme to ANVIL[19]:

– *Wave*

This is a non-editable track with the waveform of the audio of the file. It is automatically loaded when opening the video file and facilitates the annotators when they single out specific

[19] Sub-tracks are listed in brackets.

words from the transcript to associate with image parts or gestures (cf. AnchorText track explained below).

– *AvTopic*

Audiovisual Topic: this is a track in which the annotators indicate topics within the programme, taking into consideration both visual (images) and audio (sound + speech) parts of the file. In most cases, speech is indicative of the actual content/topic indication, while image (shots) indicates the exact start–end of the topic boundaries (i.e. visual change denotes the offsets). Sound change (natural sound/music, etc.) is another indication/clue of the topic offsets.

– *Transcript (Utterance, Overl-Utterance, Subtitles, OC)*

This is a group of annotation tracks that includes different types of manual transcripts. Utterance and overlapping utterance depict the transcription of the first and (whenever applicable) second-overlapping speaker (information in these tracks is directly uploaded from manual transcription files). Subtitles are a track in which the annotators can write down any subtitles that appear in the video. As subtitles we consider only textual translations of what is being spoken of, or textual translations of something written (e.g., an inscription depicted on screen, or a title page, etc.). Start time and end time of each subtitle block indicate the time-period during which the specific subtitle block is visible/present in the video. OC is a track in which the annotators can write down any—other than subtitles—text that appears visually in the video, e.g., closed-captions, labels, inscriptions that are well depicted and easy to read, etc. This means that anything an Optical Character Recogniser (OCR) running on the image could pick up is annotated in this track. Start and end times of each OC block should determine the time period during which the text is visible on screen.

– *AnchorText (Utt-Text, Overl-Utt-Text, Sub-Text, OC-Text)*

This is a group of annotation tracks in which the annotators indicate the exact token or multiword expression, which participates in a cross-media relation. It is only the head of a phrase that is being singled out to participate in a relation, making sure that modifiers and complements of this head should hold true for, e.g., the image with which the head is related. In Utt-Text, the annotator writes down the token or multiword expression that comes from an Utterance track. Similarly, the tokens that come from the Overl-Utterance track, or the Subtitles track or the OC track are written down on the corresponding Text tracks. The offsets of the token or multiword expression are denoted with the help of the waveform.

– *Images (FrameSequence, KeyframeRegion)*

This is a group of tracks in which the annotators indicate segments of the video or regions within frames of the video that participate in cross-media relations. *FrameSequence* is a segment of the video that equals a shot; it is either the background as a whole, or the foreground as a whole, that participates in the relation; so the annotators denote which part of the shot participates in the relation. In some cases, it could be both, meaning that it is what depicted as a whole that participates in the relation and not a particular segmentable region, for example, consider a sequence of frames depicting an aerial view of Athens while the speaker refers to "cities" of Northern Europe.

A *KeyframeRegion* depicts a particular object (or cluster of objects) of interest in a FrameSequence. Start and end times of such annotation block indicate the time period during which the object of interest is visible (usually the whole duration of the corresponding FrameSequence). The annotator chooses one frame from within this time period in which the object is better viewed and draws the outline of the object using the corresponding ANVIL feature [28]. The frame number is also included in the annotation details of the block.

– *HumanActivity*

This is a group of tracks in which the annotators indicate those gestures and other body movements that participate in cross-media relations. Start and end times of each of them indicate the time period during which the movement is visible, covering all phases of the movement, i.e. starting from just before the body part starts moving up until the moment when the body part is again at a rest position. The annotator indicates the body-part through which the movement is realised (hands, head, legs) and for gestures the type of gesture (deictic, iconic, emblem, metaphoric). Only those body-movements and gestures that have propositional content are annotated, unless a non-propositional gesture is considered related to something verbally expressed in a meta-information relation. For all HumanActivity tracks, the annotators indicate the human who performs the movement or gesture by creating a corresponding KeyframeRegion, drawing the outline of the human figure and including this in the relation in which the HumanActivity annotation participates.

Annotators provide also a *label/tag* that expresses what is depicted in *all human activity and image tracks*. This is used in the annotation process by the annotators as a way to check themselves that the relation they have picked up stands, indeed, between the different media.

– *Relations*

This is a group of annotations through which the different types of the COSMOROE relations are indicated. These annotation blocks do not have their own time offsets; for practical reasons, they get the ones of the AnchorText that participates in the relation. So, the annotator chooses first the AnchorText annotation block that participates in the relation, and then determines any of the following that applies: the corresponding Transcript annotation block, the participating Gesture annotation block(s), and the participating FrameSequence and KeyframeRegion annotation block(s). In case the annotators would like to denote a relation between, e.g., a Gesture and a FrameSequence (i.e. with no AnchorText participation), an empty AnchorText annotation block with the offsets that correspond to, e.g., the FrameSequence (or any offsets if this is not possible) is being created. The relations are actually non-temporal objects; it is the time offsets of the participating tracks (modalities) that define the time-extent of the relation. For some relations, the annotators must determine the sub-type of the relation and the sub-subtype of the relation. For example, for Equivalence possible subtype values are the following: token–token, type–token, metonymy and metaphor; in the case of metonymy, the annotator indicates the metonymy type too.

In developing such annotation scheme, one encounters an important issue: *which are the linguistic units, the image units and the body-movement units that should be correlated*? As evident in the above description of the COSMOROE annotation scheme, we have opted to correlate single words (or multiword expressions) with complete body-movements and (dynamic in our corpus) images of single objects, clusters of objects, foreground or background parts of images or whole images. The annotation endeavour showed that images engage in the semantic interplay with other media in many different levels of granularity, which are fully covered though with the above mentioned approach. Given the fact that COSMOROE is intended for use in computational applications, its annotation scheme makes use of image-segmentation-types used in image analysis systems (cf., for example, foreground/background distinctions).

While the annotation task focuses on cross-media relations, it is actually such that a number of by-products are also created:

- *Manual transcriptions*. Transcriptions of the speech stream of each file are created. This annotation takes place mainly at the level of a clause (neither at word-level, nor necessarily at the sentence level). Further word-level transcription takes place when particular tokens of a clause participate in a cross-media relation. Though the resulting transcription is not as granular as a completely word-level transcription, it can form a rough basis for training and/or evaluating speech recognition systems.
- *Optical characters*. All text appearing on screen, including subtitles. Such text is manually transcribed, and can be used for developing optical character recognition systems.
- *Acoustic events*. Music, ringing bells and other distinctive acoustic events in the programmes are being annotated. They can be used for developing acoustic event detection systems.
- *Audiovisual topics*. Each file is manually segmented into thematic areas (topics) according to visual and textual cues. This annotation can be used for developing audio-visual topic detection and story segmentation systems.
- *Body-movements*. Gestures and other body movements (e.g., walking) are being annotated as such and can be used for developing gesture and body movement identification systems.
- *Shots*. Frame-sequences of a single camera movement are being annotated. This annotation can be used within shot detection system development.
- *Objects and events*. Association of events mentioned in the transcript with the corresponding frame-sequences depicting the events, and association of objects mentioned in the transcript with the corresponding representative keyframeregions. Annotator-provided labels are available for these events and objects too. All such associations can be used in event and object recognition and identification system development.

Of course, in some cases, medium-specific annotations are not as granular as one may wish for the training of the corresponding systems. For example, for speech recognition, word-level transcription of the whole speech stream is ideally needed; however, this is costly and in many cases forced alignment techniques are employed for aligning manual transcriptions to their corresponding audio signals and getting a word-level transcription. It has been shown that the smaller the manual segments to be aligned are, the better results one gets from such alignment [14]. The COSMOROE-related transcription by-products could therefore assist greatly in such a task. In annotating gestures, it is the time offsets of the gesture, the participating part of the body and the type of gesture that are being annotated; what is not included is information regarding the temporal phases of the gesture (e.g., preparation phase, stroke, retraction, etc.). However, this is information may be added on the existing annotation, if needed.

3.2.2 Validation and inter-annotator agreement

As mentioned earlier, all files of the current COSMOROE annotation corpus have undergone validation. An expert annotator went over the already annotated files, correcting, adding and refining the annotations. Depending on the quality of the initial annotation, the expert annotator calculated that, on average, the task took her 60–80 times real time to

Table 5 Validation results from a 30-min COSMOROE-annotated audiovisual file

Results	Number of relations	(%) over 396 total relations
Correct relations–correct sub-types	308	77.77
Correct relations–wrong sub-sub-types	38	9.59
Wrong relations	5	1.26
Unnecessary relations	30	7.57
Missed relations	45	11.36
Total relations identified by annotator	381	

complete. This means that given the manual transcription of the speech stream and all other related transcriptions (subtitles and other optical characters present in the video), the rest of the annotation needed 1 h for 1 min of video, in the best case. The approximate time needed for both bad quality annotation (which actually needed deletion and annotation from scratch) and better quality annotation (which required mostly verification that it is correct) was within this range. We believe that this indicates that what does take time in the annotation task is the indication and identification of a relation rather than the practical part of indicating time offsets for the different entities and filling in the relation template with the participating entities.

Table 5 presents results from validating a 30-min segment of a Greek travel programme; the file was annotated after the annotator had concluded a hands-on training annotating another Greek travel programme. Validation showed that missing a relation (false negative) or indicating a relation that is not there (false positive) is more likely than choosing the wrong relation. It was observed that an important cause for all three types of mistakes was wrong segmentation of the image units, i.e. inclusion of more than one shots in the same frame-sequence (overload of visual information that could not been handled), or identification of image parts that were not clear in the information they carried (e.g., a background that is minimal because the frame is mainly covered by the foreground object; so it did not make sense to combine the background with a specific language unit, cf. Fig. 12). The latter is considered an unnecessary relation. In other cases, the annotator made assumptions on what is depicted in the image and related it to a language unit; cf. for example Fig. 14, in which the narrator says that he is on the way to Erfound, and we watch images of different landscapes. One does not know if these places are parts of the town of Erfound or other places through which the narrator travelled to reach Erfound. Drawing an equivalence (part for whole metonymy) relation between Erfound and (one or more of) these images is like forcing a relation with no evidence, so, we consider it a false positive.

This validation showed how consistent and accurate the annotators were when compared to what the expert annotator considered proper implementation of the COSMOROE framework. However, we wanted to check whether the COSMOROE relations truly reflect the semantic interplay between different media. To do so, we decided to check the agreement between different annotators. High inter-annotator agreement would give some good hints on whether the relations are descriptive and whether it is sensible for one to attempt automating the task. We calculated inter-annotator agreement on the 30-min segment of the Greek travel programme from which validation results were given before; for doing so, missing relations and unnecessary relations (as indicated during validation) are excluded. Agreement is checked on the total number of relations identified by both annotators (the original one and the expert). So, actually, this way we question the "expertise" of the validator and treat the two annotators equally, judging their agreement. We have to note that the role of expert annotators in agreement studies may take many forms (e.g., the agreement between one annotator and the majority opinion, the latter playing the role of the "expert") [10], though one should study the agreement among annotators of similar experience to draw

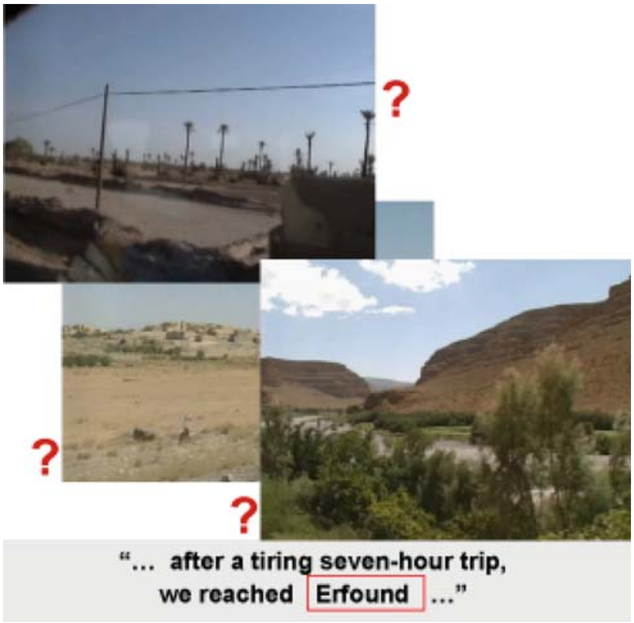


Fig. 14 Example of a false-positive relation

Table 6 Inter-annotator agreement for COSMOROE relations

	Annotator$_a$					
Annotator$_b$	**Equivalence**	Metonymy$_x$	Metonymy$_y$	**Complementarity**	**Independence**	Totals$_r$
Equivalence	236	0	0	0	2	**238**
Metonymy$_x$	0	0	19	0	0	19
Metonymy$_y$	0	19	0	0	0	19
Complementarity	0	0	0	39	0	39
Independence	3	0	0	0	33	36
Totals$_c$	**239**	**19**	**19**	**39**	**35**	**351**
Percent agreement						0.8774
Percent agreement expected						0.4901
Kappa						0.7597
Weighted kappa						0.8851

sound conclusions that can be generalised; we decided to carry out a preliminary inter-annotator agreement study for COSMOROE relations in the least favourable case, which is that between an expert and a trainee annotator.

Table 6 presents the calculations and results of the kappa statistics for inter-annotator agreement. Agreement is explored in the three main COSMOROE relations; in our data, it was noticed that apart from a few cases of disagreement with regard to the main relation itself, there was also disagreement on the sub-sub-type of Equivalence relations: it was the case of metonymies, in which, the annotators agreed on the main relation and its sub-type (i.e. Equivalence–Metonymy) but disagreed on the type of metonymy. In calculating the Kappa score, we considered this case of disagreement equally important to disagreement on the main relation; this gave us a kappa score of 0.7597. However, disagreement on the main relation, the sub-type or the sub-sub-type is of decreasing importance. If one is to reflect this in the calculation of kappa, one should calculate a weighted kappa score [17]. We used a weight of 1.00 for all cases of full agreement, a weight of 0.00 for all cases of disagreement on the main relation, a weight of 0.50 for cases of subtype disagreement (non-applicable in our case) and a weight of 0.75 for cases of sub-subtype disagreement.[20] The weights actually determine the contribution of each category to the final agreement score. The weighted kappa reached 0.8851.

3.2.3 Annotation issues and cases

Annotating a corpus of audiovisual files with cross-media relations is challenging in many respects. First of all, the task itself, i.e. to analyse the semantic relations among media for forming a message, is demanding, in terms of *cognitive effort* and *time*. This is because such analysis is multifaceted, i.e. it requires the analysis of a number of different media and modalities before one actually identifies any cross-media relations; this costs in time and increases cognitive effort for the annotator. Cognitive effort seems to be high even for indicating relations between already identified media units; as presented in the previous section, the validation of the COSMOROE annotated corpus showed that there was a number of missed relations in the files before validation. This provides evidence to the fact that the semantic interplay between media in multimedia discourse takes place unconsciously in humans, while they watch/read multimedia documents or when they are engaged in multimodal interaction themselves. They understand the meaning that emerges from cross-media interaction, but do not realise how they combine what they see with what they hear or do, while they are doing it. The case is similar to language comprehension and production, in which understanding or talking does not imply that one has knowledge or is conscious of the syntactico-semantic relations between linguistic units.

Second, using COSMOROE for the annotation requires that the annotator is/gets familiar with linguistic analysis (and corresponding terminology), since the framework draws analogies to language description and uses terms that are used mainly for describing semantic phenomena in language (cf. for example the "metonymy" relation). On top of this, there are cases in which the annotator must resolve/analyse semantic phenomena in the speech/text of the multimedia document first and then to proceed to annotating the relation between the language unit and another modality. Figure 15 illustrates such case: the word "green" is used in the speech stream metonymically, instead of the word "landscape"; it is a case where a defining property is used instead of the thing defined. One needs to be able to identify and resolve this metonymy in language, to identify and annotate the relation between the image of the green landscape and the corresponding word

[20] For details on the calculation of kappa and of the weighted kappa cf. [17].

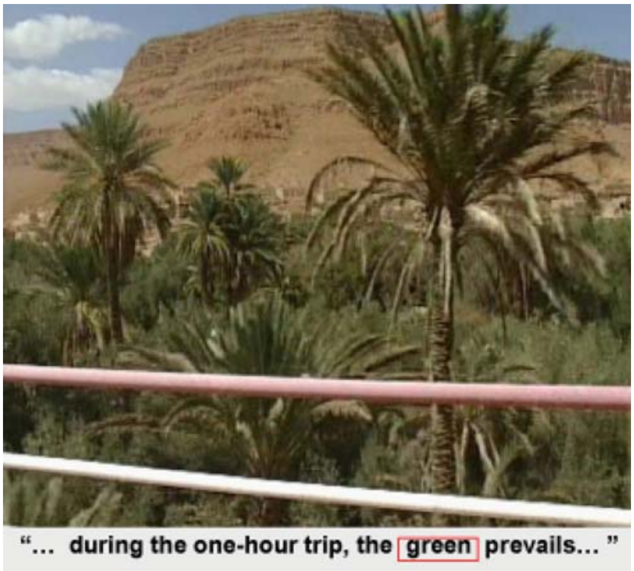


Fig. 15 Textual metonymy and equivalence relation to image

that resolves the language metonymy (i.e. landscape); in this example, the relation is a type–token one.

Furthermore, familiarity with the characteristics and types of the different media is also required by an annotator. For example, identifying the type of gesture is important, before actually the annotator decides on the cross-media relation in which the gesture is engaged. Dealing with frame-sequences and keyframes requires familiarity with image analysis concepts.

Last, in our annotation endeavour, there have been a couple of cases in which the COSMOROE relation to be chosen was not directly evident to the annotators. For example, the first time the annotators were faced with verbs like, e.g., "to enjoy oneself" they were not sure how to relate them to the corresponding video footage that depicted, e.g., someone dancing. Is this relation a metonymic one and in particular, does the image show one aspect of the concept denoted by the verb (other aspects being to, e.g., sing, etc.)? Is this relation an adjunct of manner, in which the image shows one way for one to enjoy oneself? Is this relation a type–token one, in which the verb denotes a class of actions and the image shows one instantiation (member) of such class?

An analysis of the semantics of the verb assisted in deciding on the type–token relation: the verb expresses a qualification of one's action(s), and therefore classifies the action(s) under a specific type. The qualification refers to an internal state of mind/or emotion with no clearly objective physical realisation (no facial expressions that denote positive feelings are necessarily present—one may dance because this is part of a ritual, or it is one's job, etc.; so, whether one enjoys oneself when dancing is a purely language-based characterisation of the action depicted in the video).

The case of identifying this relation as a complementary one (adjunct of manner) was ruled out, since in such case, the image should modify the action denoted by the verb providing the *means* (object) for performing the action; however, in our example, the image specifies the action by depicting a more concrete action. Last, the case of characterising the relation as metonymic (and in particular one in which one modality expresses one aspect of the concept expressed by another) was also ruled out, since no figurative equivalence is present, simply a classification of an action, according to a qualitative criterion.[21]

Such "difficult" cases in using COSMOROE for annotating a corpus of multimedia documents show the potential of such analysis of multimedia documents. It can help us reveal the characteristics of some, e.g., verbs/concepts that stem from their relation to other media, and therefore dig further into their semantics and use.

4 Discussion

COSMOROE is a "language" for describing multimedia dialectics. It is intended to capture *the interaction between pieces of information expressed through different media* in multimedia discourse. It has a twofold objective: to *deepen* our knowledge on the use of different media in discourse and to *foster* computational research on the issue (cf. also Sect. 1 on the need for the latter). COSMOROE is the only existing framework for describing the semantic relations between different media in multimedia discourse from such perspective and in this level of granularity; it draws analogies to phenomena from language discourse and focuses on the semantics that emerge from multimedia message formation. As such, it points to research questions that have not been formulated this way before and questions that can be addressed when analysing a COSMOROE annotated corpus:

- Which concepts (words) are usually visualised in accompanying images or expressed through gestures in discourse and what is their level of abstraction?
- How does metonymy or metaphor function in multimedia discourse? Does visual information provide the clues (background knowledge) that are normally needed for understanding metonymy in textual discourse?
- Which concepts are usually complemented with visual or gestural arguments, adjuncts or appositions? Could it be that one may, e.g., predict the selectional restrictions for the complements of a predicate, when knowing its visual/gestural complements (and the other way around)?

[21] Compare also the *troponymy* relation for verbs in Wordnet [22], which is a hypernym–hyponym (type–token) relation and expresses exactly this kind of argumentation that a "manner" specification is inherent in verbs that are related as hyponyms to more general verbs.

- How is exophora realised? Could one use anaphora resolution mechanisms to resolve exophora?
- What kind of meta-information is necessary in discourse and in which cases?
- In which cases "what one sees is not what one hears" in discourse? Are these cases of irony or humour?

In other words, what kind of semantic associations take place when humans combine pieces of information from different media in multimodal communication?

In using data-mining techniques in a corpus annotated with the COSMOROE relations, one will be able to address such questions, and shed some new light on how meaning emerges in multimedia discourse. Patterns of semantic interaction across media and cues that denote specific types of interaction can emerge by mining such annotated corpora.

Of course, developing algorithms for identifying the COSMOROE relations is far from trivial. In the last few years, algorithms for the computational identification of *equivalence* relations between images (or image regions) and language are being developed, in an attempt to "bridge the gap" between low-level visual information and high-level concepts/semantics, for a number of applications. The approaches are either probabilistic [3,58] or logic based [18, 47]. Learning approaches require properly annotated training corpora (such as those developed in, e.g., [19,29]) for learning the associations between images/image regions represented in feature-value vectors and corresponding textual labels, while symbolic logic approaches rely on feature-augmented ontologies [18,55]. Other approaches rely on both training corpora and ontologies [56]. These algorithms deal with cases of literal equivalence only, and in particular, simple cases of type–token associations. The limited number of annotated corpora that are available for such research consists of video keyframes and/or photograph collections enriched with a *finite set* of textual labels that have been manually associated to the images as a whole or to image regions; there is no markup of *existing associations* between modalities that are *both present* in the multimedia document collection (as done in the COSMOROE annotation endeavour).

Going beyond the automatic identification of such equivalence relations will be a bigger—though essential—challenge (cf. Sect. 1). However, up until now, any attempt to delve into computational multimedia semantics was hindered by the lack of a systematic analysis of what such semantics actually comprises of and a corresponding lack of appropriately analysed corpora; this is exactly the need that COSMOROE intends to cover.

5 Conclusion

In this paper, we presented the COSMOROE cross-media relations framework, focusing on image, language and body-movement interaction in multimedia discourse. The framework adopts a "message-formation" perspective for describing multimedia dialectics that takes into account the unique characteristics of the different media. We compared this framework with related work and implemented it in annotating a corpus of travel programmes; we argued that COSMOROE is descriptive enough to generalise over different media-pairs and that it is suitable for use in developing computational models of the corresponding multimedia semantics.

Acknowledgments The author would like to thank the annotators Maria Lada and Evi Mpandavanou for their collaboration, and Stelios Piperidis and Carl Vogel for their stimulating comments on the cross-media relations framework.

References

1. André, E., Rist, T.: The design of illustrated documents as a planning task. In: Maybury, M. (ed.) Intelligent Multimedia Interfaces, pp. 94–116, Chap. 4. AAAI Press/MIT Press, Cambridge, MA (1993)
2. André, E., Rist, T.: Referring to world objects with text and pictures. In: Proceedings of the Computational Linguistics Conference, pp. 530–534 (1994)
3. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)
4. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: a free tool for segmenting, labeling and transcribing speech. In: Proceedings of the First International Conference on Language Resources and Evaluation, pp. 1373–1376 (1998)
5. Barthes, R.: Image, Music, Text. Flamingo (1984)
6. Bateman, J., Delin, J., Allen, P.: Constraints on layout in multimodal document generation. In: Proceedings of the Workshop on Coherence in Generated Multimedia, First International Natural Language Generation Conference (2000)
7. Bateman, J., Delin, J., Henschel, R.: Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making. In: Perspectives on Multimodality, pp. 65–89. John Benjamins, Amsterdam (2004)
8. Bernsen N.: Why are analogue graphics and natural language both needed in hci? In: Paterno, F. (ed.) Interactive Systems: Design, specification and verification. Focus on Computer Graphics, pp. 235–251. Springer, Berlin (1995)
9. Bordegoni, M., Faconti, G., Feiner, S., Maybury, M., Rist, T., Ruggieri, S., Trahanias, P., Wilson, M.: A standard reference model for intelligent multimedia presentation systems. Computer Standards Interfaces **18**(6/7), 477–496 (1997)
10. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996)
11. Carlson, L., Marcu, D., Okurowski, M.: Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: Current Directions in Discourse and Dialogue, pp. 85–112. Kluwer, Dordrecht (2003)
12. de Carolis, B., Pelachaud, C., Poggi, I.: Verbal and nonverbal discourse planning, proceedings of fourth international conference on autonomous agents. In: Proceedings of the Workshop on Achieving Human-Like Behaviour in Interactive Animated Agents, Fourth International Conference on Autonomous Agents (2000)

13. Cassell, J.: A framework for gesture generation and interpretation. In: Computer Vision in Human–Machine Interaction, Chap. 11. Cambridge University Press, London (1998)
14. Chen, L., Liu, Y., Harper, M., Maia, E., McRoy, S.: Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In: Proceedings of the 4th Language Resources and Evaluation Conference (2004)
15. Corio, M., Lapalme, G.: Integrated generation of graphics and text: a corpus study. In: Proceedings of the Association of Computational Linguistics Workshop on Content Visualisation and Intermedia Representation, pp. 63–68 (1998)
16. Corio, M., Lapalme, G.: Generation of texts for information graphics. In: Proceedings of the European Workshop on Natural Languge Generation, pp. 49–58 (1999)
17. Crewson, P.: Fundamental of clinical research for radiologists: reader agreement studies. Am. J. Roentgenol. **184**, 1391–1397 (2005)
18. Dasiopoulou, S., Papastathis, V., Mezaris, V., Kompatsiaris, I., Strintzis, M.: An ontology framework for knowledge-assisted semantic video analysis and annotation. In: Proceedings of the International Workshop on Knowledge Markup and Semantic Annotation (2004)
19. Everingham, M., Gool, L.V., Williams, C., Zisserman, A.: Pascal visual object classes challenge results. World Wide Web (http://www.pascal-network.org/challenges/VOC/voc) (2005)
20. Fasciano, M., Lapalme, G.: Intentions in the co-ordinated generation of graphics and text from tabular data. Knowl. Inform. Syst. **2**(3) (2000)
21. Feiner, S., McKeown, K.: Automating the generation of co-ordinated multimedia explanations. In: Maybury, M. (ed.) Intelligent Multimedia Interfaces, pp. 117–138, chap. 5. AAAI Press/MIT Press, Cambridge, MA (1993)
22. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA (1998)
23. Green, N.: An empirical study of multimedia argumentation. In: Proceedings of the International Conference on Computational Sciences-Part I, pp. 1009–1018. Springer, Berlin (2001)
24. Gut, U., Looks, K., Thies, A., Trippel, T., Gibbon, D.: Cogest conversational gesture transcription system. Tech. rep., University of Bielefeld (2002)
25. Jackendoff, R.: Consciousness and the Computational Mind. MIT Press, Cambridge (1987)
26. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, London (2004)
27. Kipp, M.: Gesture generation by imitation—from human behavior to computer character animation. Boca Raton, Florida: Dissertation.com (2004)
28. Kipp, M.: Spatiotemporal coding in anvil. In: Proceedings of the 6th Language Resources and Evaluation Conference (2008)
29. Lin, C., Tseng, B., Smith, J.: Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. TRECVID Proceedings (2003)
30. Lindley, C., Davis, J., Nack, F., Rutledge, L.: The application of rhetorical structure theory to interactive news program generation from digital archives. Technical Report INS-R0101, Centrum voor Wiskunde en Informatica (2001)
31. Magno-Caldognetto, E., Poggio, I., Cosi, P., Cavicchio, F., Merola, G.: Multimedia score—an anvil-based annotation scheme for multimodal audio-video analysis. In: Proceedings of the LREC Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation Of Multimodal Input And Output Interfaces, pp. 29–33 (2004)
32. Mann, W., Thompson, S. : Rhetorical structure theory: description and construction of text structures. In: Kempen, G. (ed.) Natural Language Generation: New results in Artificial Intelligence, Psychology and Linguistics, pp. 85–95. Nijhoff, Dodrecht (1987)
33. Marsh, E., Domas-White, M.: A taxonomy of relationships between image and text. J. Document. **59**(6), 647–672 (2003)
34. Martin, J., Grimard, S., Alexandri, K.: On the annotation of multimodal behavior and computation of cooperation between modalities. In: Proceedings of the International Conference on Autonomous Agents workshop on Representing, Annotating, Evaluating Non-verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, pp. 1–7 (2001)
35. Martin, J., Julia, L., Cheyer, A.: A theoretical framework for multimodal user studies. In: Proceedings of the Second International Conference on Cooperative Multimodal Communication, pp. 104–110 (1998)
36. Martin, J., Kipp, M.: Annotating and measuring multimodal behaviour—tycoon metrics in the anvil tool. In: Proceedings of the Language Resources and Evaluation Conference 2002, pp. 31–35 (2002)
37. Martinec, R., Salway, A.: A system for image–text relations in new (and old) media. Vis. Commun. **4**(3), 339–374 (2005)
38. Maybury, M. (ed.): Intelligent Multimedia Interfaces. AAAI Press/MIT Press, Cambridge, MA (1993)
39. Maybury, M., Wahlster, W. (eds.): Intelligent User Interfaces. Morgan Kaufmann Publishers, San Francisco, CA (1998)
40. McNeil, D.: Gesture and Thought. The University of Chicago Press, Chicago, IL (2005)
41. Minsky, M.: The Society of Mind. Simon and Schuster Inc., NY, USA (1986)
42. Moore, J., Paris, C.: Planning text for advisory dialogues: capturing intentional and rhetorical information. Comput. Linguist. **19**(4), 651–695 (1993)
43. Moore, J., Pollack, M.: Problem for RST: the need for multi-level discourse analysis. Comput. Linguist. **18**(4), 537–544 (1992)
44. Nicholas, N.: Parameters for rhetorical structure theory ontology. In: University of Melbourne Working Papers in Linguistics, vol. 15, pp. 77–93. University of Melbourne, Melbourne (1995)
45. Pastra, K.: The language of caricature: language and drawing interaction. Final year project, Department of Greek Philology and Linguistics, University of Athens (1999) (in Greek)
46. Pastra, K.: Viewing vision–language integration as a double-grounding case. In: Proceedings of the AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research, pp. 62–67 (2004)
47. Pastra, K.: Vision–language integration: a double-grounding case. Ph.D. thesis, University of Sheffield (2005)
48. Pastra, K.: Beyond multimedia integration: corpora and annotations for cross-media decision mechanisms. In: Proceedings of the 5th Language Resources and Evaluation Conference, pp. 499–504 (2006)
49. Pastra, K., Piperidis, S.: Video search: new challenges in the pervasive digital video era. J. Virtual Reality Broadcast. **3**(11) (2006)
50. Pastra, K., Saggion, H., Wilks, Y.: Intelligent indexing of crime-scene photographs. IEEE Intell. Syst. **18**(1), 55–61 (2003)
51. Pastra, K., Wilks, Y.: Vision–language integration in AI: a reality check. In: Proceedings of the 16th European Conference in Artificial Intelligence, pp. 937–941 (2004)
52. Radev, D.: A common theory of information fusion from multiple text sources. step one: cross document structure. In: Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue, pp. 74–83 (2000)
53. Rocchi, C., Zancanaro, M.: Generation of video documentaries from discourse structures. In: Proceedings of the 9th European Workshop on Natural Language Generation (EWNLG 9) (2003)

54. Sanders, T., Spooren, W., Noordman, L.: Toward a taxonomy of coherence relations. Discourse Process. **15**, 1–35 (1992)
55. Simou, N., Tzouvaras, V., Avrithis, Y., Stamou, G., Kollias, S.: A visual descriptor ontology for multimedia reasoning. In: Proceedings of the workshop on Image Analysis for Multimedia Interactive Services (WIAMIS) (2005)
56. Srikanth, M., Varner, J., Bowden, M., Moldovan, D.: Exploiting ontologies for authomatic image annotation. In: Proceedings of the ACM Special Interest Group in Information Retrieval (SIGIR), pp. 552–558 (2005)
57. Taboada, M., Mann, W.: Rhetorical structure theory: looking back and moving ahead. Discourse Stud. **8**(3), 423–459 (2006)
58. Wachsmuth, S., Stevenson, S., Dickinson, S.: Towards a framework for learning structured shape models from text-annotated images. In: Proceedings of the HLT-NAACL Workshop on Learning Word Meaning from non-linguistic Data (2003)
59. Whittaker, S., Walker, M.: Toward a theory of multi-modal interaction. In: Proceedings of the National Conference on Artificial Intelligence Workshop on Multi-modal Interaction (1991)