# COSMOROE Annotation Guide - Cross-media semantic relations in multisensory and multimodal discourse

**Technical Report** · February 2015

CITATIONS
0

READS
281

**1 author:**

Katerina Pastra
Athena-Research and Innovation Center in Information, Communication and Knowledge Technologies
**44** PUBLICATIONS **386** CITATIONS

SEE PROFILE


# TECHNICAL REPORT SERIES

# COSMOROE ANNOTATION GUIDE

Cognitive Systems Research Institute
ISSN: 2407-9952

This document was prepared by:


KATERINA PASTRA

Cognitive Systems Research Institute (CSRI)
7 Makedonomachou Prantouna Street
15125, Athens, Greece
kpastra@csri.gr

Acknowledgements:

Eirini Balta (COSMOROE Search Engine Development)

Panagiotis Dimitrakis (COSMOROE Annotation Processing and Automatic Validation)

Annotation Team:

Maria Lada, Evi Mpandavanou, Argiro Vatakis, Maria Koutsombogera, Elina Desypri, Niky Efthymiou and Aggeliki Altani

Research related to this work has been supported by the

FP7 Project POETICON++ (Grant No 288382)

FP7 Project POETICON (Grant No 215843) and a

John Latsis Foundation Grant.

To be cited as:

Katerina Pastra (2015). COSMOROE Annotation Guide. CSRI Technical Report Series, CSRI-TRS-150201, ISSN 2407-9952, Cognitive Systems Research Institute, Athens, Greece.

This document is available from: www.csri.gr/technical-reports

Any request shall be addressed to kpastra@csri.gr

## Table of Contents

## Abstract

This technical report provides guidance on annotating semantic relations mainly between language, images and sounds as they occur in naturalistic contexts, such as audiovisual material and follow the COSMOROE cross-media semantics framework.

COSMOROE (CMR) is a descriptive framework for modeling the semantic interplay between different means of expression, when formulating multimodal messages. It identifies a number of semantic association types through which integration of modalities is served in multimodal message formation processes (Pastra 2008). In this document, we present an annotation scheme for COSMOROE-based analysis of multimedia documents of any kind. The annotation required is multi-faceted and with a number of by-products. Thus, this technical report aims at providing guidelines for use by any annotator regardless scientific background and expertise. To this end, examples have been included, as well as tips and lists of 'interesting' cases as we have indicated them in the last 5 years, through the annotation of TV travel series, newspaper caricatures and Hollywood movies by several teams of annotators with diverse backgrounds. In what follows, one focuses on annotation of audiovisual (video) documents; however, annotation of static images (captioned or surrounded by accompanying text) or other files with any combination of language, image, and sound follows along the same lines.

Examples of annotated data are available to download from the CSRI Website (Downloads Section) and can also be accessed through the COSMOROE Search Engine at: http://www.cosmoroe.eu

# COSMOROE Annotation Guide

Behaviour understanding and generation require that humans employ fundamental cognitive mechanisms and modules in a dynamic, distributed and thus, highly interactive way. Language, Perception and the Motor System are engaged into such interaction along with –among others- prior generalized knowledge of the world (semantic memory), and strong inferential mechanisms (reasoning). Cutting edge experimental research in Neuroscience and Cognitive Psychology provides evidence of a tight integration between these modules and sheds light on the fundamental mechanisms employed for achieving it.

Natural recordings or simulations of everyday interaction and behaviours of any kind invoke such integration during information processing. These recordings may take the form of e.g., photographs, videos, films, graphics, and others. Some of these remain purely naturalistic, others are more artistic, and in some cases the former mix with the latter. In these creations, Language (text, speech), Images (static, moving, 2D, 3D, of entities and/or movements/human activity), and Sounds (acoustic events, music) interact in meaningful ways formulating messages. The more artistic the genre through which the message is built, the more eclectic this semantic interplay is.

There is a vast amount of multimedia data, created by professionals or laymen and their sheer production is increasing rapidly: TV productions, illustrated documents (such as newspapers, books, blogs, and encyclopedias), captioned photo albums (in social media or within official archives, e.g., in crime scene investigation), homemade videos, surveillance videos, education or cultural heritage related audiovisual archives, verbally or gesturally commanded video games are just some examples. As we process such messages, we employ our cognitive system to trace this integration for making sense out of it, predicting and interpreting continuously as the message (or its processing) evolves dynamically in time.

However, what is it that we trace though? In other words, what do we see as we listen, or what do we read as we see? How is speech/text associated to accompanying images/video of objects and actions and corresponding sounds? Understanding semantic association processes in integrating language, images, and sounds can contribute radically in employing critical thinking both when processing information created by others and generating audiovisual messages ourselves.

COSMOROE (CMR) is a descriptive framework for modeling the semantic interplay between different means of expression, when formulating multimodal messages. It identifies a number of semantic association types through which integration of modalities is served in multimodal message formation processes (Pastra 2008). In this document, we present an annotation scheme for COSMOROE-based analysis of multimedia documents of any kind. The annotation required is multi-faceted and with a number of by-products.

Thus, this technical report aims at providing guidelines for use by any annotator regardless scientific background and expertise. To this end, examples have been included, as well as tips and lists of 'interesting' cases as we have indicated them in the last 5 years, through the annotation of TV travel series, newspaper caricatures and Hollywood movies by several teams of annotators with diverse backgrounds. In what follows, one focuses on annotation of audiovisual (video) documents; however, annotation of static images (captioned or surrounded by accompanying text) or other files with any combination of language, image, and sound follows along the same lines.

Examples of annotated data are available to download from the CSRI Website (Downloads Section) and can also be accessed through the COSMOROE Search Engine at: http://www.cosmoroe.eu

## COSMOROE Annotation Objectives

The main objective of a COSMOROE annotation session is the indication of semantic relations among individual language, image, and sound units (any combination of them). The annotation comprises indication of a relation type from the COSMOROE set and indication of relation arguments, through annotation of their time offsets (i.e.,, start-end time) and/or position (spatial reference). In particular, segmentation of both the audio and video streams is needed:

a) Audio segmentation: segmentation of speech into utterances & speaker turns, identification of acoustic events, music-no music segments, and tokens of interest per utterance (word, multiword expressions, head-only of a phrase). The latter are only those language units that will be used as arguments in semantic relations.

b) Video segmentation: segmentation of the video into shots and indication of regions of interest (ROIs) per shot; a region of interest may be a particular keyframe region (the contour of which one may draw on the keyframe), the foreground of a shot or shot segment, the background of a shot or shot segment, the whole shot itself, or a shot segment in which both foreground and background are of interest. Video segments that show a gesture are indicated by type; for each gesture and body movement, the visual regions of interest depicting the agent, the tool and affected object (if applicable) of the action are also annotated and linked to them. All visual units (images of objects/scenes, gestures, body movements) are tagged without listening to the audio or reading any accompanying text. Tagging is considered a verbal categorization process. Last, text appearing in the video stream (in the form of subtitles, closed captions, graphic text or scene text) is also transcribed and segmented into units of interest (as with speech).

Identification of language or video/image units of interest for use as relation arguments goes part and parcel with the process of identifying a semantic relation and indicating its type. Thus, processing-wise, one performs segmentation of the audio into utterances and speaker turns, transcription of the utterances, segmentation of the video into shots,

segmentation and transcription of all graphic and scene text and then one proceeds to identification of relations and candidate arguments. When relation arguments have been decided, annotation of these arguments takes place (segmentation, and indication of their types etc.) as well as indication of the relation type. When all possible relations have been identified, labeling of all visual arguments should take place without interference from accompanying language. In substituting the visual arguments of a relation with their tags, the annotator may get a list of verbally expressed relation triplets in order to check the validity of the identified relations in the conceptual space.

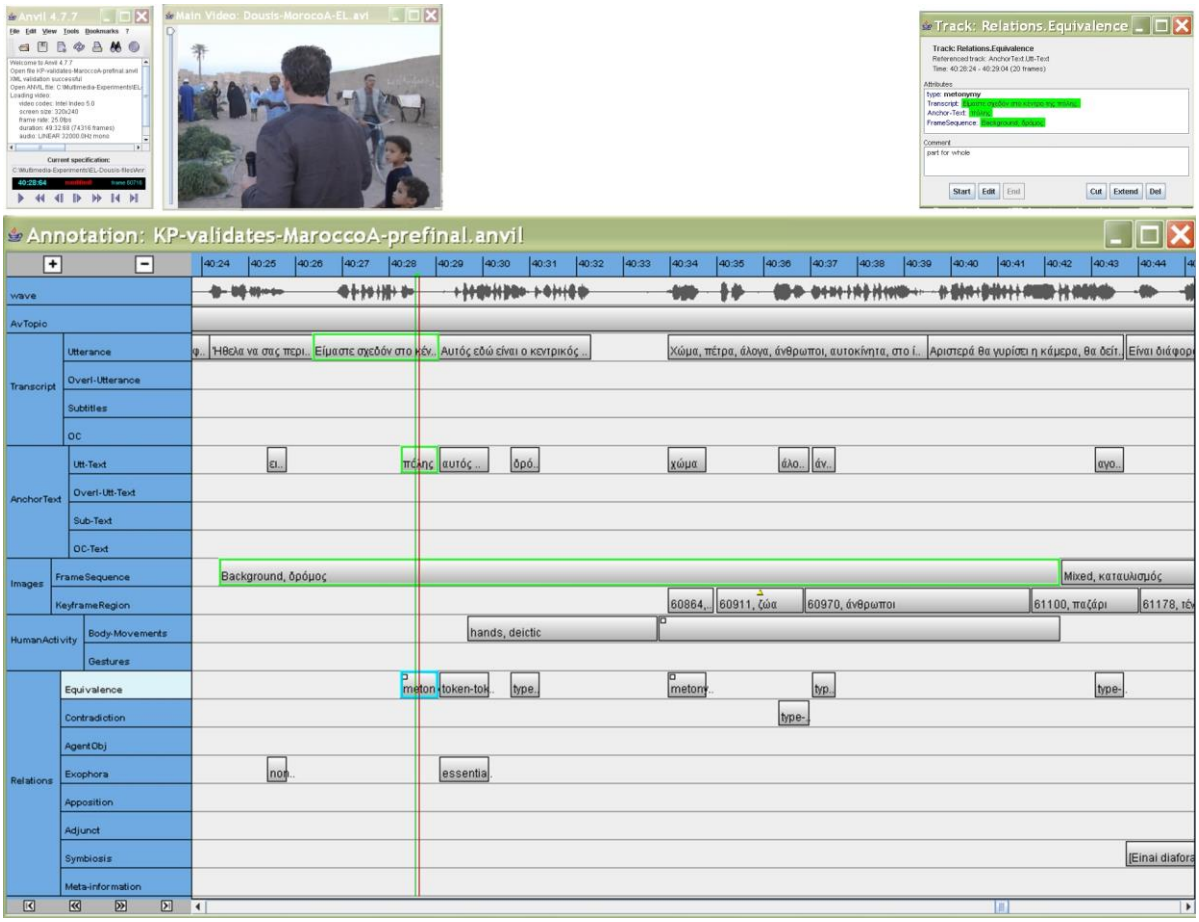


**Figure 1: An annotation session in the ANVIL annotation environment.**

We will go through the details of this process, following the annotation specification file available at http://www.cosmoroe.eu.

## Annotation Tools

One may use an audio annotation environment, such as the Transcriber Tool (Barras et al. 2000) for speech transcription at the utterance level, speaker turn annotation and acoustic event identification in the audio stream of the audiovisual file. For all video stream related annotation, and the annotation of semantic relations, one may use an audiovisual annotation environment such as the ANVIL tool (Kipp 2012). Within such environment, one loads the specification file, the audiovisual file and the audio stream

annotation (all of it, or just those tracks that are of interest for the current annotation session, i.e.,, the transcript and the acoustic events).

## Annotation Tracks

The specification file comprises the following annotation tracks:

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

**Audiovisual Topic** (AvTopic): this is a track in which the annotators indicate subtopics within the TV travel documentary, 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 images (shots) indicate 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.

**Acoustic Event** (AcousticEvent): this is a track in which the annotators indicate acoustic events within the TV travel series. These are non-speech sounds such as the barking of a dog, the horn of a carriage, the steps of a person on a ladder and so on (see Annex I). Note that sounds are generated by **actions; there is no sound without an underlying action.** In that sense, acoustic events are the acoustic representation of action-related concepts. In CMR annotation, these sounds may be correlated with the action that generates them (and which is visually or verbally expressed), or may stand for a denoted action and function complimentary to something shown or said. The annotation of acoustic events comprises annotation of their time offsets and may take place in a different annotation environment (e.g., Transcriber) while speech transcription takes place. This is preferable so that the annotator singles out acoustic events based on the acoustic signal only (i.e., with no access to the corresponding visual stream that may affect detection of such events); in the audiovisual environment the identification (labeling) of such events may be facilitated using information from the visual stream too.

**Transcript** ***(Utterance, Overl-Utterance, Subtitles, OC)***: This is a group of annotation tracks that includes different types of transcripts. **Utterance** is the track in which speech transcription performed with the Transcriber is loaded. Each utterance is visualized in the annotation environment as one block with time offsets as indicated in the transcription file. **Overlapping-Utterance** is the track in which the transcription of overlapping speech sections is loaded, as indicated in the transcription file. **Subtitles** is 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., of graphic text). 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. **Optical Characters (OC)** is a track in which the annotators can write down any -other than subtitles- text that appears visually in the video, e.g., close-captions, labels, shop signs that are well depicted and easy to read etc. In other words, in this track, one annotates anything that an Optical Character Recognition System (OCR) running on the image could pick up. Start and end times of each OC block should determine the time period during which the text is visible on screen. The ***OC-Graphic Text*** is all text that appears on screen by the producers (e.g., close captions). ***OC-Scene Text*** is all text that appears in the natural scenes depicted on screen e.g., shop signs, car plates etc.

**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 multi-word expression which participates in a cross-media relation. This is the language unit that participates in the relation. Segmentation of the utterance into units that participate in the semantic interplay with images should follow simple principles: a complete meaning should be encoded in the unit, with no modifications or complements that are not essential for the generalized meaning of the concept denoted by the unit (e.g., 'car' rather than 'my car', 'plays basketball' etc.). In Utt-Text, the annotator captures the token or multi-word expression that comes from an Utterance track. Similarly for those that come from the Overl-Utterance track, the Subtitles track or the OC track. The offsets of the token or multi-word expression are denoted with the help of the waveform.

Note: phrases should be only multiword expressions/terms, i.e., combinations of tokens that stand as one referent. In other words, one should prefer the use of the headword of a phrase unless the headword does not stand on its own in discourse (e.g., "green house" is a term, one should not use just the token "house" for denoting the AnchorText).

**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 (mixed), 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 Southern Europe. Generally speaking, it is the camera movement, angle and filming effects that guide one to decide whether the foreground, the background or the whole/mixed shot is of interest.

***Foreground vs. Background*** distinction tips:

- Foreground is usually whatever stands closer to the camera and is more clearly seen (zoomed in). Sometimes, foreground is moving and background is static, but not always:

  Close Static – Distant static --> foreground is whatever is closer to the viewer;

  Close Moving – Distant moving --> one may see e.g., a presenter walking on the pavement and cars moving in the road- whatever is closer to the camera is the foreground;

  Close Static – Distant moving --> one may see e.g., a standing presenter zoomed in and moving cars behind him (traffic); the presenter is the foreground, the cars are the background;

  Close Moving – Distant static --> one may see e.g., people walking in a haste in the town; the moving entities are the foreground, the static scene is the background.

- In some shots, the camera starts with an overview of a place and then zooms into something specific; this shot is mixed and when tagged, it is the zoomed object that is to be tagged. In the comments field of this annotation track, extra tags may be added, related to the overview part too.

- Usually zoomed in: foreground - Usually zoomed out: background

**KeyframeRegion** depicts a particular object of interest in a FrameSequence/shot. Start and end times of such annotation blocks indicate the time period during which the object of interest is visible. The annotator chooses one frame from within this time period in which the object is better viewed and includes the frame number in the annotation details of the block. Annotators can also draw the contour of the object (with as many markers as needed to denote the shape of the object). Alternatively, to the interest of time, bounding boxes around the object may be drawn too.

Tips for object contour drawing: the objective is to draw a contour that is representative of the shape of the object. To this effect:

- we create rounded contours (by drawing a very dense polygon) when needed;

- the contour should have a single start and finish (can be thought of as a continues line drawing; e.g., the contour of a glass should start from a point like the place where our lips touch which will allow one to draw the whole contour at once);

- in case of occlusions we draw the contour in a way that will serve the drawing objective (even if one has to follow the inferred contour of the object).

One should be careful to draw either on the original size of the video (i.e., the frame size that appears when you open the video file) or on a customized size of known percentage to

the original. This is because the object drawing information in the ANVIL annotation environment is a list of the x and y position of each marker/bullet one has drawn, i.e., the connected bullets that form the object outline. The start of the x and y axis (i.e., the 0,0 position) is the upper left point of the frame. Now, the x and y values make sense, when one knows the frame size. We only know the original frame size of the video and not the size of the video after one has customized it, unless one has used a standard percentage from the View menu.

Some tips on shot segmentation (frameSequence segmentation) and keyframe Region annotation:

- Fade in - Fade out parts of a shot are excluded from the annotation block completely, or they are divided between subsequent shots according to what is more intelligible.
- More than one FrameSequence tracks may be available in the specification file to accommodate for cases when we want to comment on both e.g., the background and the foreground of a shot (they participate in different relations) and for cases of split screen (i.e., screen split in two or more windows).
- More than one KeyframeRegion tracks are available in the specification file to accommodate for annotation of more than one objects at a time.
- Keyframe Reference number: do not use the last or first frame, just the best, the one in which the object is best viewed.
- Use the FrameSequence Foreground when you want to refer to a cluster of entities rather than one or two specific objects in the foreground (e.g., crowd, traffic etc.).
- In Keyframe regions one draws only SINGLE objects.
- Single object in the Foreground can be annotated either as FrameSequence Foreground or as Keyframe Region; if it is a human figure any of these choices is fine. For other objects, we prefer the use of a Keyframe Region annotation (so contour information is provided too).
- When something is largely occluded, one should avoid annotating it, e.g., black taxi (in the background) hardly seen due to occlusion by e.g., foreground objects covering most of the frame, such as the image of the narrator talking.
- For a contour to be representative of the object shape, the contour has to be precise.
- Objects whose contour does not denote the distinctive shape of the object (due to occlusions, or because the object is too small, or it is not shown from a good angle etc.) should not been annotated.

In the label field of all images, a tag is assigned that should actually categorize what is depicted (cf. section "Visual Labeling").

**Human Activity *(Body Movements, Gestures)***: This is a group of tracks in which annotators indicate visually perceived human activity in the form of gestures and body movements that participate in cross-media relations. Start and end time of each body movement or gesture block should 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. For both gestures and body movements, their exact type is also indicated (i.e., deictic, iconic, emblem, metaphoric – cf. Annex II). Only those body-movements and gestures that have propositional content are annotated. The effector (body part) used in the body-movement or gesture is indicated by selecting a value from a predefined drop down list. Body movements may also be used for annotating animal activity.

Whenever a Body Movement or Gesture annotation takes place, corresponding keyframeRegion annotations of the agent, the artifact extending the effector (tool) and the affected object (if applicable) are also provided. If possible, the agent figure should be annotated in a frame that depicts the peak of the movement, i.e., a frame which though static is representative of the movement undertaken; the time-offsets of the agent figure annotation are those that denote the start-end of the figure appearing in a frame sequence. Similarly, tools and affected objects are annotated in keyframes in which they are clearly visible. The specification file allows for direct linking of movements and their agent/tool/affected object keyframeRegion annotations. In the label field, the annotators tag the Body-Movement or gesture (cf. section: Visual labeling), through a label that categorizes the movement in terms of its goal.

Sometimes, one needs to annotate a state rather than an actual movement (e.g., someone laid down, someone standing etc.). In such cases, one uses the Body Movement track, denoting though that it is the whole posture that is involved and providing a label that comprises a past participle.

Additional Body Movement and Gesture tracks are available to accommodate movements that take place parallel in time (i.e., in multi-person interaction).

**Relations *(Equivalence, Exophora, Apposition, Adjunct, Symbiosis, Contradiction, Meta-Information)***: This is a group of tracks through which the different types of COSMOROE cross-media interaction relations are indicated. Depending on the annotation environment to be used, these relations a) may inherit the time offsets of the language unit that participates in the relation (in ANVIL specification file), or b) may have their own time offsets that start as soon as their earlier in time argument starts and finish when their latest in time argument finishes, as in the ELAN tool (Wittenburg et al., 2006) specification file.

In either case, this is a practical issue, related to the structure of the specification file that is best served in the different annotation environments. A relation is defined by its type and arguments. So, the annotator chooses image (or sound) and language units from the annotation tracks mentioned above.

For some relations, the annotators must determine the sub-type of the relation. For example, for the Equivalence Relation, possible subtype values are: token-token, token-type, metonymy, metaphor. The Metonymy Relation has further subtypes, which are also to be indicated; the direction of the relation is indicated as well, though in most cases, there direction can be inferred automatically by looking at the relation type. For example, in the vast majority of token-type relations, the token is the visual argument.

Note: No utterance should be left completely unassociated; by definition, visual units and language are in a symbiosis relation, when no other semantic association can be drawn.

## Relation Types (definitions and examples)

The relation types to be used in annotating the semantic interplay between language and vision correspond to three major interaction relations, each one with its own subtypes: Equivalence, Complementarity and Independence.

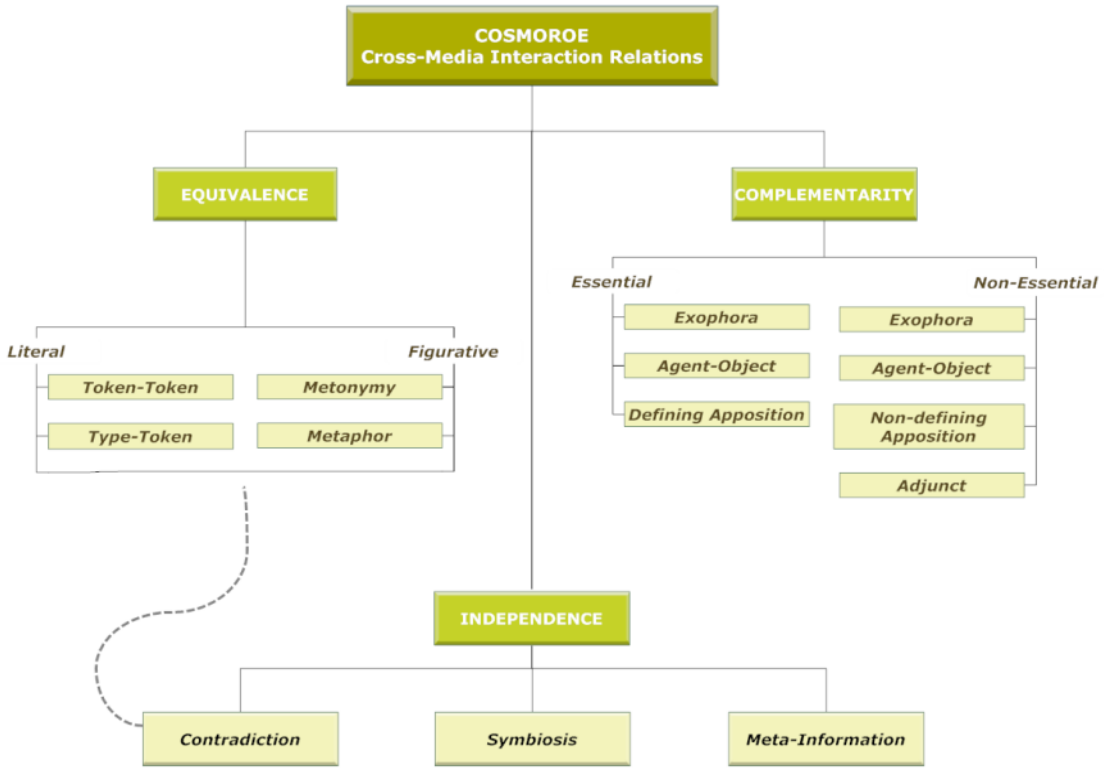


Figure 2: The COSMOROE Releation Set.

**Equivalence** (Multimedia Message comprises X and Y, and X = Y): the information expressed by the different modalities is semantically equivalent; it refers to the same entity (object), action, feature etc. Drawing an analogy to language discourse, equivalence relations could be thought of as paradigmatic relations between modalities. There are four subtypes of semantic equivalence: ***token-token, token-type, metonymy and metaphor***; the first two denote ***literal equivalence*** and the other two denote ***figurative equivalence.***

In detail:

**Token-token**: in such cases, both language and image refer to exactly the same entity, uniquely identified as such. 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 two modalities.

**Token-type**: in these cases, language expresses the category of the entity instance (instances) shown in the image. Reference to e.g., "helmets" while showing someone wearing a helmet is a token-type relation case, in which an object category is instantiated with a specific type of helmets (it could be any other type of helmets depicted). The object category linked with a visual through such IsA relation may be of *any level of specificity*. Figure 3 presents an example in which the word "housing" and images of several types of houses are related through such token-type equivalence relation. The token-type relations could be thought of as instructions to an algorithm that the modality denoting the "type" (category) may have a number of tokens (instantiations) and actually the more general the category denoted is, the more instantiations it will actually have.


**Figure 3: A token-type relation example, between the word "housing" and images of blocks of flats and other types of accommodation.**

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:

**Metonymy**: the two referents come from the same domain, they have similar associations, there is no transfer of qualities from one referent to another. For example, someone talks of the notion of "monarchy" and shows an image of a crown, or someone talks about the "US department of Defense" and an image of the Pentagon is presented at the same time, someone talks about the "US president" and the viewer watches an image of the White House. 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, or one is a material the other is a thing made of this material. Such cases are sometimes considered to be cases of Synecdoche; we will not differentiate these cases, we will consider all of them to be cases of metonymy. As an example of metonymy consider figure 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.

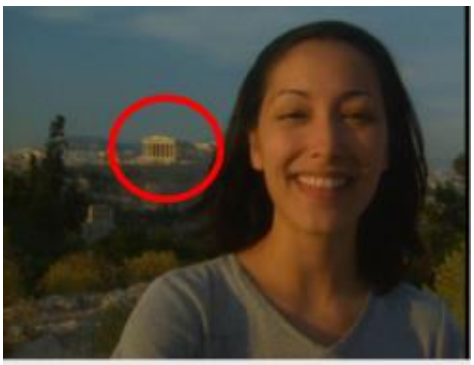


**Figure 4: A metonymy example, between the word "Athens" and the image of the Acropolis.**

***Note***: There are many different metonymic patterns. The annotator should indicate the metonymic pattern by making a selection from a dedicated drop-down list. See Annex III for a detailed list.

In many cases, the audiovisual message comprises also language or even visual metonymies, i.e., there is a metonymy in the utterance or image itself. The annotator should solve this metonymy first, and then decide on the semantic relation between language and image. In other words, the annotator must indicate the modality-specific phenomenon, and resolve it. One may use a pattern of the form: phenomenon:type:uttText:resolution, e.g., [metonymy:entity for its feature:green:landscape] in which the word "green" refers to the "green landscape". Such patterns may be written in the 'comments field' of each annotated element. The element can participate in a relation, but, the annotator should use the resolved reference of the element as argument of that relation, i.e., in this example, it is the word "landscape" that should be correlated to the corresponding image (with a *token-type* relation).

In other cases, there are visual metonymies, i.e., the image of an object stands for something else (is a symbol of something else) and this something else is related to the utterance. For example, consider the case in which you have the utterance: "Mauritius has a long history" and the image of a house. The image of the house is a visual metonymy (*part for whole*) for the image of a residential area. The denoted image of a

residential area stands metonymically as part of Mauritius. The annotators proceed as mentioned above with the visual metonymies too, i.e., they write down the phenomenon pattern at the comments section of the relation. The resolution of the phenomenon should satisfy the relation.

All metonymic patterns used are expressed with an expected Image to Language direction e.g., tool for action (image shows the tool, language refers to the action). However, most patterns may be employed with the opposite direction too e.g., Language to Image. Thus, metonymy direction is explicitly denoted by the annotators and when the metonymy in the comments field also needs to have a different direction than the default, the annotator may add a flag (i) to denote the inversion of the default direction, e.g., Metonymy:part for whole:town:region (i).

The flag denotes that the metonymic pattern should be inverted, i.e., that the "town" denotes the whole.

**Metaphor**: one draws a similarity between two referents which actually belong to different domains; there is a transfer of qualities from one to another. For example, someone says "the giant is here" and the image shows a big man. The "giant" is intended to mean the specific man who is big like a giant, and not any giant literally. Figure 5 illustrates another 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 comprises 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).


**Figure 5: A metaphor relation between the word "serene" and a gesture.**

**Complementarity** (Multimedia Message comprises X + [(Y)]): the information expressed through one modality is (an essential or optional) complement of the information expressed in another. Association signals (e.g., verbal indexicals pointing to an image or image part) indicate cases of essential Complementarity, while non-essential Complementarity is characterized by one modality 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 verbally). Complementarity relations have a syntagmatic (syntactic) nature.

Complementarity sub-relations are clustered into two groups: those in which complementarity between the pieces of information expressed by each modality 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 modalities is combined. Explicit or implicit cues must be present in discourse for one to characterize 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 modality. In the case of non-essential complementarity, one modality provides extra information to what the other expresses, information that is not vital for the comprehension of the latter. In particular:

**Essential Exophora**: These are cases of anaphora, in which signals of semantic equivalence 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 modality in which they are being expressed (e.g., speech) and another modality (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 that, 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, cf. for example figure 6, in which the deictic word "this" signals that somewhere in the context (image region highlighted in red) one will find its reference. The signals itself is semantically "empty". The most frequent equivalence signals are indexical words, deictic words, and deictic gestures which point to an image or image part.

Figure 6: An essential exophora relation between the word "this" (in "the pollution has taken its toll on this.") and the image region showing the Parthenon.

**Non-essential Exophora**: These are cases of anaphora in which the entity referred to through one modality is revealed in another, though this is not vital for understanding the message. Figure 7 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 nationalized 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 nominalized adjectives does not complicate or hinder communication.


Figure 7: A non-essential exophora example.

**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). In this relation, phenomena of ellipsis 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 some languages is the subject of a predicate totally missing. Information on the subject may be evasive (cf. for example passive voice impersonal constructions) but still present. However, consider cases of de-verbal nouns, 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 information. Figure 8 illustrates one such case.

**Figure 8: Essential agent-object relation example.**

**Non-essential Agent-Object** This is a case of one modality providing information on the missing agent or object of an action/state/event expressed by another, though not vital for understanding the message. In such cases, the missing agent or object are known from the wider communication context or from the shortly preceding multimedia discourse or they are intentionally left vague. For example, consider figure 9, in which the travel documentary presenter says "we went shopping in Oxford Street" while the video shows images of clothing 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).

**Figure 9: A non-essential agent-object relation example.**

**Defining Apposition** In defining apposition, one modality 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 e.g., see figure 10, "The president of Greece" and the image of Konstantinos Stefanopoulos. What makes this situation different from the *token-type equivalence* relation is that it is tied

**Figure 10: A defining apposition example.**

to the specific context and should not be considered as generally valid (i.e., the word president is not used only for the specific person, but for a number of other people too and Mr. Stefanopoulos himself was/will not always be a president). It is not the same case as in e.g., the association of the word "furniture" with the photograph of a chair, which though crossing over different conceptual levels, it is not tied to a specific context of discourse.

**Non-Defining Apposition**: In this case, one modality reveals a generic property/characteristic of the very concrete entity mentioned by another. For example, someone may say 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 modality. 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 visualize the shape of something or the colour/hue etc.). In non-defining apposition, we focus on cases in which the extra information provided by the visual modality allows attribution of a quality, physical characteristic or role to an entity (e.g., occupation of a person, ethnic origin etc.).

***Adjunct*** This relation denotes an adverbial-type modification (place-position, means, source). One (or more) modalities function as adjuncts to the information carried by another. In figure 11, the presenter of the travel documentary mentions that she is "heading to" an island, while the corresponding image/video shows a high speed ferry boat; the image reveals the means used to visit the island, it actually complements the predicate "to head to a place".

Figure 11: An adjunct relation example.

**Independence** (Multimedia Message comprises X, Y): In this third interaction relation type, each modality curries 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 multimodal message (it is like a conjunction of sentences). The relation of independence comprises three subtypes:

**Contradiction** Usually in artistic genres of 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; cf. 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 token-type contradiction. A token-token contradiction could be one between the word "Einstein" and the image of one's five year old daughter. 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 multimodal message may be characterized 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 mistakes in creating a multimodal document, such as mistakes of synchronization between the speech/audio and the images depicted at the same time, or human mistakes in describing entities/situations. Contradictions are present in professional tv programmes too; in such cases they may also indicate mistakes in video post-editing (see for example figure 12).

**Figure 12: A token-type contradiction example**

**Symbiosis**: Symbiosis is a case of different pieces of information being expressed by the modalities, 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 the sake of creating a multimodal message. Cf. for example the case of someone saying: "Democracy was nurtured during the Golden age of Pericles" and the corresponding images zooming into the narrator. In this case, the image is a visual filler, just showing the speaker, but nothing related to what is being said. TV talk shows, news programmes etc. are full of such cases.

**Meta-information**: Last, this is a case in which one modality reveals extra information through its specific means of realization (or through specific non propositional types of it); it forms part of the multimodal message, and due to its nature (non-propositional) it

stands independently but inherently related to the information expressed by the other modality. For example, consider part of a TV travel documentary in which the narrator mentions e.g., that she is "traveling 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 multimodal 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 "traveling", (since the filming/the camera is traveling - 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 multimodal 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 multimodal message, they carry extra information independent from pieces of information expressed by the other modalities, 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 modalities).

### Comments in Relations - Resolution of Semantic Phenomena

In the "comment" field of the relations we note and resolve different semantic phenomena that appear in either the language or the visual elements, such as metonymies, metaphors, paraphrases, word sense clarifications, antonomasia etc. For example, one may want to denote that "wander around with the car" means to "drive around", i.e., one may provide a paraphrase that enables elaboration of the corresponding CMR relation drawn between the language unit and the visual one. To do so, we follow the pattern suggested for metonymies (cf. section Metonymy); phenomenon:subtype:utt-text:resolution e.g., paraphrase:null:wander around with the car:drive around.

If more than one such comment is needed in the same relation, we write each one in a new line. Also, series of such resolutions may be needed for one element, i.e., the depth of the resolution steps maybe more than 1. In that case, we write the different, sequential patterns in new lines.

Comments that justify the relation, should be added in a Symbiosis relation in special cases e.g., when it is due to poor image quality, or ambiguity on whether what is being said is what one sees.

Special case: In the *entity for feature* metonymy subtype, one adds in the comments field the type of feature denoted, i.e.,

feature:feature-type:anchortext:iconic gesture tag e.g.,

feature:shape:table:square

## Visual Labeling

In the label field of all visual elements, the annotator notes a tag that categorizes what is depicted. This has many uses, one of which is for the annotators to check themselves that the relation they have picked up stands, indeed, conceptually between the different modalities. These tags should be single words or multi-word expressions, literal, and one should avoid making inferences on what is depicted but rather stay as close as possible to what is actually shown, regardless context. The latter implies that visual labeling should better be performed as an individual process before or after all relational annotation and without watching the full video or listening in parallel to the audio stream of the video.

- Gestures: name the gesture that is performed (propositional content – only for iconic, metaphoric gestures and emblems);
- Body Movements: use a verb to name the goal of the movement performed, e.g., “cut”; use a past participle to describe a state e.g., laid down;
- Keyframe Region: name the object depicted e.g., “man”;
- FrameSequence-Foreground/Background: name what is depicted only in the foreground/background, e.g., “buildings”;
- FrameSequence-Mixed: give a label that characterizes the totality of what is depicted, e.g., “city”.

Tags such as: town, traffic, road etc. are generic in nature and could be further detailed in the "comments" section, using a comma separated list of more specific things that are shown. For example, the tag "town" can be further elaborated through this list of tags: “buildings, pavement, road, vehicles, people” (if shown).

Tags like "shop"/"store" could be further elaborated using a word denoting the products sold in the store, e.g., "clothes" (unless the original label was "clothes shop"). Other cases: “shop window” --> “shoes”, “pavement” --> “road”, “vehicles” etc.

A translation of the visual labels may also be provided in a dedicated field.

## Tips

- do not annotate video segments that are repeated in exactly the same way during the file e.g., the logo of the show and corresponding OC text;
- do not include visual arguments in a relation, if they are far away before or after the language argument, and in the meantime other utterances have also been uttered;

- do not make assumptions in relations - we relate only what we are sure of; for example, the narrator may say that he is on the way to Erfound, and one watches 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 traveled 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;

- in cases of "subject - link verb – attribute" syntactic patterns, do not correlate both the subject and its attribute with the same image (just the subject), e.g., "Mauritius is a beautiful island" - there is no need to correlate both Mauritius + image (of the island) and "island" + image of the island. Just do the former.

- Reference Resolution task: if one is interested in using the annotated corpus for a reference resolution task, one should:

  - use speaker change information from the transcription file, and

  - use the existing annotation of essential or non-essential Exophora which provides pronominal resolution for personal and possessive pronouns when language provides anchors for hooking the parallel speaker image info.

NOTE: pronouns are resolved once per utterance (in case of repeated pronoun in the utterance, the same resolution is implied) – a personal pronoun denoting a speaker is always linked to the corresponding image of the speaker when the speaker appears for the first time. From then on, the relation is not drawn again, because it can be easily inferred from this initial link and the speaker turn annotation (see also 'clusters of annotation cases' below).

## Clusters of annotation cases

### Acoustic Events in CMR relations

In most cases, an acoustic event (e.g., sound of a man running) engages into a *token-token* relation with the visual depiction of the event (e.g., man running), and in a *token-type* relation with the verbal expression of the event (e.g., "man runs"). This is because the non-speech auditory modality and the visual modality denote –by definition- instances (tokens) unless they are symbolic. On the other hand, language denotes categories (types) unless expressing unique entities (e.g., named entities) or being non-directly referential (e.g., in the case of deictics).

### Co-reference Resolution/Speaker Identification

One should make sure that each speaker is annotated once in the file, i.e., a speaker name (if given) and the image of the speaker are linked through a token-token equivalence

relation. Similarly, the personal pronoun denoting a speaker should be linked to the corresponding image of the speaker when the speaker talks for the first time. Subsequent relations of this type are only optional.

Change of speaker is usually denoted through an Exophora relation, linking the image of the new speaker and the personal pronoun used in his/her utterance (e.g., "I cannot answer that": essential exophora relation between "I" and the image of the speaker); however, there are cases in which no AnchorText is present to anchor such Exophora relation (e.g., "cannot answer that" – cf. pro-drop languages in which the use of the personal pronoun is not necessary, since it is the inflection of the verb that denotes the person information). In such cases one may use the AgentObject relation.

Related to the above, one resolves a pronominal co-reference/anaphora in every utterance once, i.e., if a pronoun is repeated in the utterance we do not resolve it through an exophora relation again. Also, we resolve a pronominal coreference/anaphora to indicate that the speaker has changed. We resolve a pronominal coreference/anaphora when no other relation takes place in an utterance (to avoid symbiosis).

- When verbs in second person singular or plural denote the audience, there is no relation to be drawn (unless we see the audience).

### Deictics

Deictic Gestures stand in an equivalence Token-Token relation with corresponding textual deictics, whenever present; i.e., they have the same referent, so they are equivalent – to the extent this is possible given their unique modality-specific nature.

Deictics (linguistic or gestural) get their value/reference (are resolved) from the corresponding Image (ImageRegion or FrameSequence). This is a case of essential Exophora.

Note 1: An arrow depicted graphically or a road sign in the form of an arrow is a Deictic Visual Symbol (image).

Note 2: When a deictic language unit is in a token-token relation with a deictic gesture, it is enough to link one of these to the visual reference with the essential exophora relation. Transitivity allows easy inference of the same relation to the other deictic unit too.

Note 3: The reference Visual Unit must have some overlap in time with the deictic unit (direct reference), since by nature they are tightly coupled; it may start earlier or/and finish later than the deictic. We are interested in what one can see as the deictic is carried out. In some cases, due to the fact that the camera does not follow what the speaker says very well, the image of the entity pointed to by the deictic may precede or follow – but these are exceptions.

Other deictic cases:

- Relative location expressed verbally (up, down etc.) + deictic gesture ☐ they are related through a *token-token* relation; the image of the place they point to resolves their reference.

- A Body Movement may be used as deictic and therefore it may behave as a deictic gesture e.g., instead of pointing with a finger, one may point to something with an umbrella; this is a body movement (not a classical deictic gesture) which is though deictic and participates in corresponding relations with what is being said and shown.

- Use of lexical deictic units as modifiers e.g., "this man", may correspond in time with Body Movements, such as embracing of the man, grasping, touching him etc. In such cases, the goal of the body movement is deictic, i.e., to show beyond doubt who the referent of the linguistic unit "this" is. Thus, the deictic word can be associated to the Body Movement, as if the latter was a deictic gesture. The word "man" is associated through a *token-type relation* to the image of the man (which is also the affected object of the body movement).

- A deictic reference e.g., "***here***" that is normally resolved with a location image reference, may be resolved with a human image reference in case the deictic word is not used as an adverbial, as in: "my man here".

### Emotions

Verbal expressions of emotions (e.g., "joy") may co-occur with a body movement (e.g., dancing) or facial expression (e.g., smiling); in such cases, the two modalities are related through an *aspect for concept* metonymy. The action/facial expression grounds the meaning of the emotion illustrating one of its aspects.

### Geographic Terms

Region, capital, island, city, river, mountain…

All such terms denote entities that have geographical boundaries on a map. So, they have a *token-type* relation with their visualisation on a map, or a *token-token* relation with their visualisation on a map if named (e.g., the region Maroussi). Usually they are engaged in metonymic relations with images (e.g., part for whole ones, the image of a beach showing part of an island).

Sea, ocean…

These terms denote entities with no strict geographical boundaries – open ended- with no parts with visual variation either, as in different parts of a city/capital etc. So, if named, e.g., Indian Ocean, they are linked with the corresponding visual through a *token-token* relation (not a *part for whole* metonymy); when not named they engage into a *token-type* relation with the image of a sea.

Bottom of the sea, landscape, place, market…

These entities are usually in a *token-type* relation with their corresponding images, unless named (e.g., Port Luis market: *token-token*). For the "market", it is far-fetched to have a part for whole metonymy, because ONE image is representative of the type (other images would just show more stalls, the pattern is the same). Whereas in a city, there is visual variation in different neighbourhoods etc. (we even divide cities in smaller regions, we do not do the same with markets...). This is for OPEN markets – if the "market" word sense is more general "the market of a town" (including e.g., shops of any kind) then we DO have a *part for whole* metonymy.

### Greetings

Body movements used when greeting someone (e.g., handshake) + stereotypical greeting expressions (e.g., "hi") are linked through *action for goal* metonymies.

Greeting Gestures (emblems) (e.g., for "hi") + stereotypical greeting expressions (e.g., "hi") are linked through token-token relations.

### Institutions

- sense 1

building (image) - ***institution*** (word): in this case we have a *token-type* relation, e.g., school building – "school".

- sense 2

people (image) - ***institution*** (word): in this case we have an *aspect for abstract entity* metoynymy, e.g., priest – "church".

- sense 3

building (image) - **institution** (word): in this case we have an *aspect for abstract entity* metonymy, e.g., church building – "religion" (or the word church itself but referring to the religion in general), school building – "education" (or the word school itself but referring to education as in "the school shapes children's personality…").

### Qualifying nouns

We refer to nouns denoting:

- Nationality / Ethnicity (Greeks, Africans etc.)
- Religion (Muslims, Christians etc.)
- Occupation (musician, doctor etc.)
- Age & Gender (old man, child, girl etc.)

- Social dimension/role (proletariat, inhabitant, immigrant, local, tourist, owner, landlady etc.)

- Situation specific role (passenger, guest etc.)

- Kinship (mother, grandfather, friend etc.)

Such nouns qualify the human entity they refer to and the qualifications are not always visually verifiable. In both cases, the language unit is analysed as a textual metonymy case of *entity for feature* (metonymy:entity for feature:owner:man) and then an *apposition* relation is drawn between the language unit and the corresponding visual. When the qualification is visually verifiable (e.g., African, doctor), we have a case of non-defining apposition, whereas when the qualification is not visually verifiable (e.g., owner), the apposition is defining. In some cases (e.g., musician, passenger), the qualification is related to an action, e.g., a musician is someone who plays music, a passenger is someone who is being transported with a vehicle; in such cases, the visual to be linked with the language unit is the visual of the characteristic action the human performs.

- nouns denoting abstract concepts that qualify life

"poverty", "richness", "luxury" etc.: these are abstract feature concepts, aspects of which are illustrated in the images, i.e., they engage into *aspect for abstract feature* metonymic relations with the image. They qualify the concrete entities or actions depicted visually; the latter ground these abstract concepts visually.

**Several words denoting: Buildings – Natural Bodies – Notion of Life**

Annotations related to *shops/any building*

The relation between the word "shop" or similar and any representative view of it e.g., en face main window/vitrine with view of entrance and optionally good view inside from entrance is *token-type*; just entrance: *part for whole*; inside only: *part for whole*; focus on specific thing/object sold in the shop: *object for action*.

Annotations related to the *underground*

The term could be used to refer to (a) the train or (b) the underground area. Option (a) creates a *token-type* relation with the image of the train, while option (b) creates a *part for whole* metonymy with images showing the platform, the tube, escalators etc.

Annotations related to the notion of l*ife*

"Life"/"live" – person/animal (image): abstract feature-aspect metonymy

"Life"/"live" – domiciles (image); there is a visual metonymy in the image (location for entity: domiciles – people); when solved, the "life" – people relation is an abstract feature-aspect *metonymy* case.

"Life"/"live" - images of activities of animate beings: abstract feature-aspect relation; (the sense "to live a life" is referred to here, not the sense "to exist")

Annotations related to unique *natural bodies*

Natural bodies, e.g., "the sun" are unique and in that sense their verbal reference engages into *token-token* rather than *token-type* relations with their corresponding images.

### Specific words with image-defined reference value

"Image" (e.g., "incredible images"): in *essential exophora* relation with the co-occurring in time visual units;

"colours" (e.g., "beautiful colours"): in *essential exophora* relation with the co-occurring in time visual units;

"view" (e.g., "great view"): ditto;

"thing"/"element"/"feature" (e.g., "it's a nice thing"): ditto;

"experience" (e.g., "You should live the experience…"): ditto;

"moments": if no complement present, we do not relate it to anything; if there is a complement (e.g., moments of happiness) we relate its complement to the image, if possible.

### Attention Verbs

e.g: "look" (with no object) + Image of something one should look at: these two engage into an *AgentObject Complimentarity relation*.

### Verbs expressing goal

"to play" – Images of Body Movements: jumping, pushing; verbal and visual units engage into an *action for goal metonymy* relation in this case. Language expresses the goal of a number of concrete actions.

But, consider another case too:

"play" – Images of Body Movements: playing football, hide and seek, etc.; verbal and visual units in this case engage into a *token-type* relation. Language expresses a category of games, such as football etc.

"to work" – Image of Body Movement: fishing; verbal and visual units engage into an *action for goal* relation.

"to prepare something" (e.g., "to prepare dinner") – Images of Body Movements: boiling, baking etc.; verbal and visual units engage into an *action for goal* relation (to prepare dinner is the final goal of the actions).

"to wait for" – Images of Body Movements: sitting down, standing, walking; verbal and visual units engage into an *action for goal* relation.

"to rest" – Images of Body Movements/States: sitting down, reading, listening to music etc. Verbal and visual units engage into an *action for goal* relation.

"to enjoy oneself" – Images of Body Movements: singing, dancing, eating, drinking etc. Verbal and visual units engage into an *action for goal* relation.

### Verbs expressing states

"to stand" – Image of someone standing: *token-type relation*

"laid down" – Image of someone lying down: *token-type* relation (the participle denotes the "result"/end phase of the body movement denoted by the corresponding verb), but:

"to lie down" – Image of someone changing stance from e.g., standing to lying down: *token-type relation;* Note: if the image shows someone lying down, then the relation is a *result for action* one.

### Verbs denoting temporal phases (aspect)

It is the complement of such verbs which denote a temporal phase (start, end, continue, etc.) that may participate in a CMR relation.

### Verbs with inherent perspective

Verbs denoting movement in general AND having a perspective (e.g.," go" vs. "come" – same movement reference different perspective regarding whether one goes away from the speaker or comes close to the speaker), are treated as *defining apposition* cases rather than *token-type* cases with the corresponding Body Movements, e.g., "come" – Image of someone walking.

### Trigger action for action

This metonymy type is frequent with perception verbs/actions and the corresponding visual or acoustic representation of other actions that trigger the perception, e.g., "to hear" – sound of e.g., a phone ringing.

### Natural Force/Phenomenon

In many cases, phenomena such as 'raining', 'snowing' etc. are depicted and/or talked about or actions that do not denote human or animal activity but rather something done by a natural force e.g., 'air shutting a window'. Such cases should be treated in a new annotation track with the name 'Natural Force Activity'; the track will be similar in structure to the Body Movement track, however, no effector field will be attributed. The agent is the natural force itself, there may be a tool used (e.g., tornado "using" a tree as a tool to affect a car), and there may be an affected object too. For phenomena, no such complements will be annotated (i.e., the rain has no tool and affected object – it has a location that could be mentioned though, and which is the place it falls at).

### Visual Symbols

Bullets on maps that denote e.g., countries, cities etc. are visual symbols. They engage into a pure visual metonymy (i.e., the bullet stands for a city); such bullet and the name of the city/country it stands for are linked through a *token-token* relation. Flags are also visual symbols of countries; the image of a flag and the name of the country it stands for are linked through a *token-token* relation (in the comments field of the visual element, one should denote that it is a visual symbol).

## Annotation Post Processing

The annotation process is labour and time intensive; it has been estimated that audio annotation left aside, all other parts of the annotation take on average 40xReal Time to complete, depending on the familiarity of the annotator with the process and how rich the file is in semantic relations (estimations were based on annotation of TV travel documentaries which are very dense with image-language relations, and the performance of an annotator who is familiar with the scheme).

Thus, once the annotation process has been concluded, the annotated files should go through a series of automatic consistency checks and a conceptual validation phase to ensure that mistakes have been avoided; consistency checking and conceptual validation scripts running on ANVIL xml COSMOROE annotation files have been implemented in PERL and are available for download at the CSRI webpage (Downloads Section).

### Consistency Checking

Consistency checking aims at identifying mistakes in data entering, omissions, incompatibilities and so on. We have identified a total of 25 checks involving data entry (e.g., only digits as frame numbers, drawing of objects within video bounds, correct insertion of elements in tracks), relation creation (e.g., both arguments of a relation should be included) and completeness of argument and relation attributes (e.g., metonymy relations have metonymy direction, keyframeRegions should have a tag etc.).

### Conceptual Validation

For each relation in an annotation file, a list of triplets is extracted in the form: Visual Element Tag – RELATION TYPE – Language Element Tag, or
Visual Tag – RELATION TYPE – Visual Tag
(the latter, when both arguments of the relation are visual units). The list of triplets is used by the annotator to check the conceptual validity of the annotated relations, and make corrections when needed.

### COSMOROE Search Engine

After validation, the annotation files are further processed for extraction of relations and their arguments and splitting of the original audiovisual file into video segments and keyframes corresponding to the visual arguments of these relations. This is the input to a search engine that has been developed for showing image-language associations from COSMOROE annotated files (see Pastra and Balta 2009) and fully updated version live at: http://www.cosmoroe.eu

Figure 13 shows an overview of the complete annotation process:

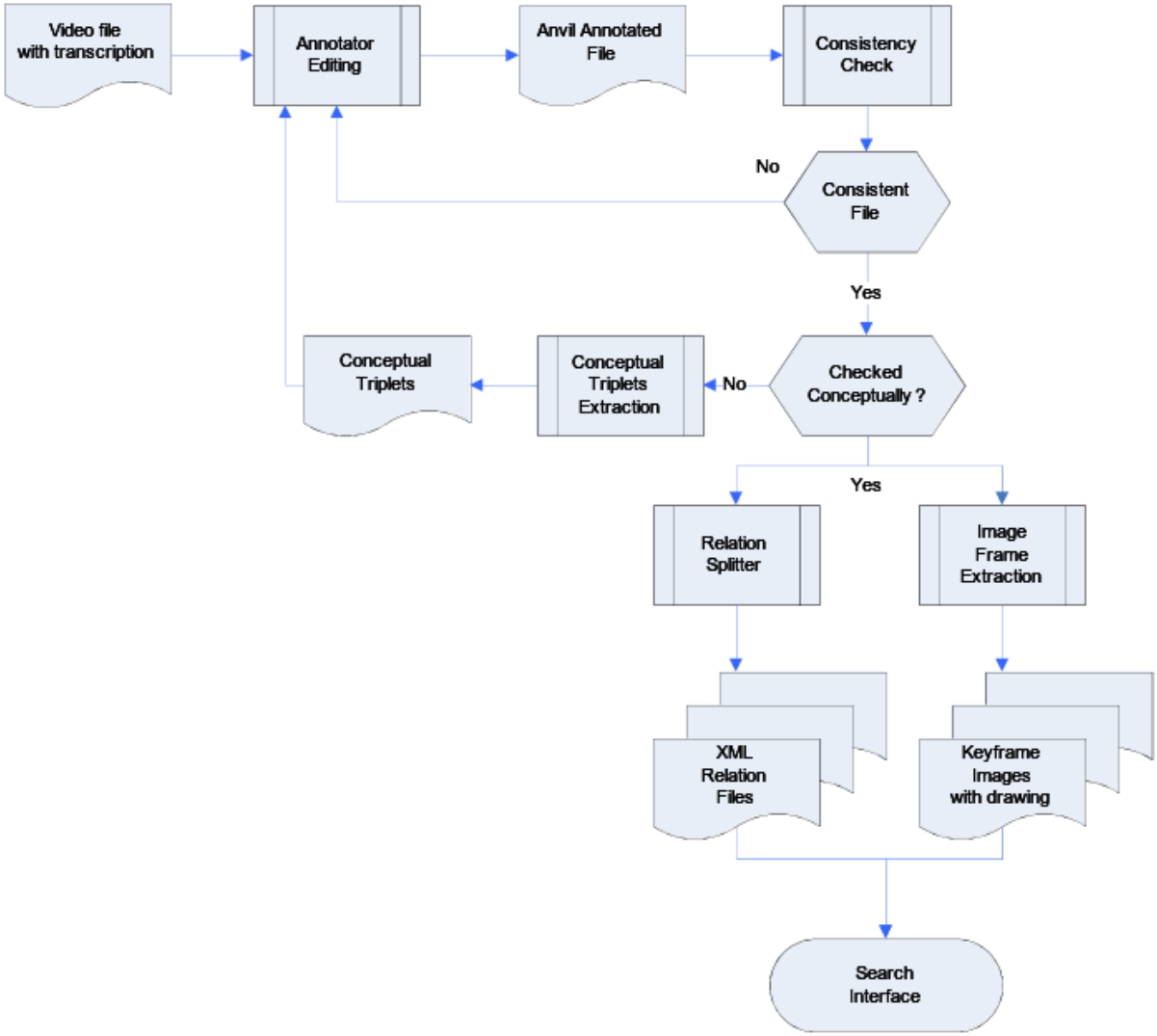


**Figure 13: COSMOROE annotation workflow; from transcription to validation and presentation.**

## References


Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2000). Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication, 33, 1-2.

Kipp, M. (2012). Multimedia Annotation, Querying and Analysis in ANVIL. In: M. Maybury (ed.) Multimedia Information Extraction, Chapter 19, Wiley - IEEE Computer Society Press.

Pastra K. (2008). COSMOROE: A Cross-Media Relations Framework for Modelling Multimedia Dialectics", Multimedia Systems Journal, vol 14(5), pp. 299-323, Springer Verlag.

Pastra K. and Aloimonos Y. (2012). The Minimalist Grammar of Action. Philosophical Transactions of the Royal Society B, 367(1585):103.

Pastra K. and Balta E. (2009). A Text-Based Search Interface for Multimedia Dialectics. In: Proceedings of the System Demonstration Session of the 12th Conference of the European Association for Computational Linguistics, pp. 53-56, Athens, Greece.

Vatakis A., Pastra K. and Dimitrakis P. (2014), Acquisition of object knowledge through Exploratory Acts, 15th International Multisensory Research Forum (IMRF), 11-14 June, Amsterdam, NL.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

## Annex I: Acoustic Events

What follows is a list of acoustic events that have been identified in different CMR annotation files. TV travel series have few acoustic events, mostly general ones (sound of a crowd, traffic etc.) and rarely do these engage into CMR relations; however, in a different genre, that of films/movies there is a big variety of acoustic events. We provide a compilation that comprises abbreviated form and definition. We also list some pronunciation-related acoustic events (e.g. whispering), which are not meant to be employed in CMR relations, however, one may annotate them in the auditory stream for other purposes.

### *Specific Acoustic Events*

- [applause]-applause: striking the palms of the hands together repeatedly.
- [baa]-sheep: sheep vocalization.
- [back cn]-back crowd noise: the sound of the murmur of a crowd in the background.
- [back conv]-background conversation: talk between two or more people in the auditory background.
- [back cs]-back crowd shouting: the sound of loud, inarticulate shouting or loud cries expressing strong emotions of a crowd in the background.
- [back explosion]-background explosion: background violent shattering caused by a bomb.
- [back laugh]-background laugh
- [back singing]-background singing: background musical sounds with the voice, especially words with a set tune.
- [band]-band: the music sound coming from a band (music band, marching band etc.).
- [baraag]-elephant: the vocalization of an elephant, trumpeting.
- [battle]: the sound of armed men fighting fiercely (shouting, dashing, clashing their weapons-swords etc.)
- [bc]-baby crying: the sound of a baby/child who sheds tears.
- [bell]-bell: tolling of a bell.
- [bells]-bells: the sound of bells tolling.
- [bump]-bump: sound of a hard hit.
- [buzz]-buzz: make a high-pitched whistling or buzzing sound.
- [cackle]-bird: the sound of cackling made by a hen or a rooster (cluck).
- [carriage]-carriage: the sound produced by a four-wheeled passenger vehicle pulled by horses while it is in movement.
- [caw]-bird: the harsh cry/the sound of a crow or a raven.
- [chain]:the sound of chains/metallic sound.
- [cheep]-bird sound: tweet, warble, bird vocalization of small bird/-s.
- [cheering]-cheering: shout for joy/celebration or in praise or encouragement.

- [chimes]-clock chimes: a melodious ringing sound produced by a clock to indicate the time.
- [chirr]-insect: the short vibrant or trilled sound, characteristic of an insect (as a grasshopper or cicada).
- [choke]: the sound made by someone being choked.
- [chorus]-chorus: a group of people performing together a song, words or tunes by making musical sounds with their voice.
- [clap]-clap: striking the palms of the hands together once.
- [clink]-glass/metal struck: a sharp ringing sound, such as that made by striking metal or glass.
- [cough]-cough: to expel air from the lungs with a sudden sharp sound.
- [cn]-crowd noise: the sound of the murmur of a crowd.
- [crackle]-crackle: the sound of a rusty gate that opens, or a wooden floor when someone steps on it, or a wooden chair that goes back and forth.
- [creak]: the sound of a rusty gate hinge, like a harsh cry.
- [crying]-crying: the sound of a person who sheds tears, a loud inarticulate scream expressing a powerful feeling or emotion.
- [cs]-crowd shouting: the sound of loud, inarticulate shouting or loud cries expressing strong emotions of a crowd.
- [dog barking]-dog barking: a sharp explosive cry of a dog.
- [drums]-drums: a rapid succession of beats sounded on a drum, often used to introduce an announcement or an event.
- [door]-door: the sound of opening and shutting the door.
- [door key]: the sound made by the keys while locking or unlocking the door.
- [e]-expiration:  exhalation.
- [explosion]-explosion: a violent shattering caused by a bomb.
- [flames]-flame sound: the sound of burning wood or other materials.
- [flip]: the sound produced while turning the page of a book etc., turn over with a sudden quick movement.
- [flutter]-bird sound: fluttering of the wings, to flap wings rapidly.
- [infant talk]-human sound: vocal sound produced by a human infant, unintelligible sounds, nonsense talk by infants.
- [gaggle]-bird sound: the cackle of a geese.
- [gallop]-gallop: the sound of galloping hooves.
- [glass breaking]-glass breaking: a sharp cracking sound/ a loud snap produced when the glass is separated into pieces.
- [gobble]-bird: bird vocalization, typically a turkey.
- [groan]-pain sound: a loud deep sound of grief or pain.
- [hit]: the sound of a punch or a hit during a fight.
- [hoof]: clatter of hooves, a continuous rattling sound of hooves (when the horse is pacing).
- [horn]-horn: the sound of a car horn.
- [horse snort]: a sudden explosive sound through the nose, especially when excited or frightened.
- [knock]-knock: a sudden short sound caused by a blow on a door.
- [augh]-laugh of the speaker.

- [liquid]-liquid: the sound of liquid fall in a small stream (trickle)/ like pouring water.
- [marching]: the sound of walking in a military manner with a regular measured tread.
- [moo]-cow: deep resonant vocal sound of cattle.
- [monkey]: monkey vocalization.
- [neigh]-neigh: high whinnying sound made by a horse.
- [phew]-human sound: used to express disgust at or as if at an unpleasant odor.
- [pulley]: the sound of a metallic wheel, metallic squeaks of a pulley
- [r]-respiration: breathing.
- [rain]-rain: a repetitive pattering sound of rain drops.
- [ring]: the sound of a phone or bell ringing.
- [roar]: a full, deep, prolonged cry uttered by a lion or other large wild animal.
- [rooster call]-bird sound: the sound of the call of a rooster, usually in the morning.
- [running]: the sound of footsteps at a speed faster than walk.
- [scr]-scratch: to tear or mark a surface with something sharp or jagged.
- [scream]-scream: a long, loud, high pitched piercing cry of fear.
- [screech]-engine sound: the high pitched sound the tires of a car make when it turns at high speed, or the high pitched sound the rails of a train make at the subway at a very high speed.
- [sh]-shouting: a loud, sharp utterance - a loud cry expressing a strong emotion or calling attention.
- [shing]: sound of a sword drawn from a sheath.
- [shots]-weapon: the sound of explosive impact (weapon, gun).
- [sigh]-sigh: the sound of a deep breath.
- [singing]-singing: musical sounds with the voice, especially words with a set tune.
- [sip]-sip: the sound (sucking sound) while drinking.
- [siren]-siren/ambulance: a loud prolonged signal or warning sound.
- [skate]-skate: the sound of the wheels of a skateboard on the ground.
- [slam]-slam: a loud bang caused by the forceful shutting of something, such as a door.
- [sobs]-sobs: crying noisily making convulsive gasps.
- [spear]: spear fighting / a clapping, rattling sound of clashing spears.
- [splat]: landing with a smacking sound.
- [splash]-water sound: to dash water upon, to wade or agitate in water, to make a splashing sound in water.
- [slap]-slap: the sound of a stroke or a blow with the palm of the hand.
- [steps]-steps: the sound of footsteps.
- [swash]: to dash violently, make a noise of clashing swords.
- [tap]-tap: the sound of striking lightly.
- [traffic]-traffic: the sound/noise the vehicles make when they move on a highway.
- [train]-train: the sound of train horn.
- [thunder]-thunder: a loud rumbling or crashing noise after a lightning flash.
- [trumpet]-trumpet: the sound made by the brass musical instrument with a flared bell and a bright, penetrating tone.

- [w] - walking: the sound of walking on the ground with soft treads at a regular pace.
- [wave]-wave: the water sound of the waves at the sea.
- [wind]-wind: the sound of the natural movement of the air/ air blowing.
- [whip]-whip: a sharp blow or stroke with a whip or lash.
- [whistle]-whistle: a clear, high-pitched sound made by forcing breath through a small hole between partly closed lips or between one's teeth.
- [whu]-owl: the vocalization of an owl.
- [ws]-water sound: the sound of water flow.
- [zing]: a shrill humming sound of metal, while cutting (with a machine) something metallic (i.e., iron).
- [zip]: the sound of a pulling a zip (to close or open something e.g. purse, trousers etc.).

*General Acoustic Events*

- [as]-animal sound: any noise of animals which cannot be described in a more specific way.
- [bird sound]-bird sound: any sound made by bird(s) which cannot be defined in a more specific way.
- [es]-engine sound: any engine sound which cannot be described in a more specific way.
- [hs]-human sound: any sound made by humans (voice) which cannot be defined in a more specific way.
- [ns]-natural sound: any noise from the environment which cannot be described in a more specific way.
- [ps]-pain sound: any sound made by humans expressing pain which cannot be defined in a more specific way.

*Acoustic events with no parallel speech and with a considerable duration that makes them stand as independent auditory segments*

[GAP (BACK SINGING)]
[GAP (BC)]-GAP baby crying
[GAP (CARRIAGE)] – GAP carriage moving
[GAP (CHORUS)] – GAP chorus singing
[GAP (CONV)]- GAP background conversation
[GAP (CN)]-GAP crowd noise
[GAP (CRYING)]-GAP crying
[GAP (CS)]-GAP crowd shouting
[GAP (FLAMES)]- GAP flame sound
[GAP (NS)]-GAP natural sound
[GAP (SINGING)]
[GAP (SOBS)]

*Pronunciation events*

- [pron=bsh]-pronunciation = background shouting: a loud, sharp utterance in the background.
- [pron=cr]-pronunciation = crying: the utterance of a person who sheds tears.
- [pron=faint]-pronunciation = faint/weak: a weak/faint utterance.
- [pron=sh]-pronunciation = shouting: a loud, sharp utterance.
- [pron=unintel]-pronunciation = unintelligible: an utterance which is difficult to understand.
- [pron=whi]-pronunciation = whisper: the utterance of a person who whispers.

## Annex II: Gesture types & Body Movements

Some gestures are phatic, non-deliberate/symptomatic, or expressive of a psychological state; they are not part of an utterance in the sense that they do not have propositional content. In other cases, some gestures seem to point somewhere, but actually they point nowhere, or they point to conceptual/mental space expressing an attempt to structure the speaker's thought. Others are discourse ones, i.e., they are associated with what is being said, they carry meaning. These are the gestures we are interested in, for COSMOROE annotation. In what follows, we provide a compilation of gesture types drawing from a wide range of disciplines that explore gestures from different perspectives:

**Emblem**: These are symbolic gestures that are consciously produced and which are usually culture-specific. For example, consider the gesture for denoting 'ok'. They usually participate in Token-type relations.

**Deictic**: Pointing gestures. They usually participate in essential exophora relations with something said or shown, which gives them semantic value. They also engage into token-token relations with verbal deictics.

**Metaphoric**: These are gestures that represent abstract concepts. Their form comes from a common metaphor e.g., the gesture for "on and on"; the concept represented through the gesture has no physical form. Subtypes: process-metaphoric (the information is depicted as an object e.g., gesture to denote "this part of the talk"), metaphoric pointing gesture (gesture for associating features with people e.g., someone pointing to person_X saying "semantics research"). The form of such gestures varies a lot. Language is needed to understand what they mean in discourse. Metaphoric gestures give shape to something abstract, they are unplanned, spontaneous (emblems are planned); it is difficult to say in advance what they stand for exactly. Metaphoric gestures usually participate in Equivalence-Metaphor relations with what is being said.

**Iconic – feature pantomime**: These are gestures that display object features (e.g., shape, size), motoric features (e.g., spatial trajectory) or spatial relations. For example, while uttering "I can throw it and it will make small jumps in the air", one may enact the "small jumps in the air" with her hand with or without the actual object present. Iconic gestures usually participate in metonymic relations with language, of the type "defining property for thing defined by the property" (e.g., language unit: "table" – visual unit: iconic gesture of something square).

**Iconic – action pantomime**: these are gestures that enact actions. In general, they enact what is being said. They are like pantomime. In the normal cases, if an action is enacted, the tool and/or the affected object of the action are NOT present in the enactment, but they participate virtually (e.g., in the enactment of *writing with a pen,* the hand is configured as if holding a pen while moving to write).

**Iconic – pantomime – metaphoric**: same as above, BUT, when the action is enacted, the hand is used to substitute the tool and/or affected object (e.g., the hand is configured to simulate the pair of scissors in the enactment of *cutting with a pair of scissors*, i.e., as if the hand is the pair of scissors).

**Beats** = these are gestures that are physically oriented to an interlocutor. They play a role in regulating the interaction with others, the transitions in discourse etc. They do not have propositional content, but they may provide "meta-information" in discourse. We normally do not annotate these.

We distinguish body movements into:

**Goal-Directed**: these are body movements that take place deliberately by the agent for attaining a goal. They comprise both transitive and intransitive actions (e.g., grasping a spoon and running respectively).

**Exploratory acts**: this is a subtype of goal-directed body movements, i.e., ones that are used for object exploration. These include the following manipulations: *Holding* (Turning vs. Static), *Picking Up*, *Putting Down*, *Rubbing*, *Contour Following*, *Touching*, *Pressing*, and *Tapping* (Vatakis et al. 2014).

**Unintentional**: these are body movements that take place involuntarily, such as *falling down*, i.e., they are effects of some cause beyond the agent's intentions.

**Demonstration**: this is enactment of the use of an object or something that happens to the object, with both tools and affected objects present. So, real enactment of an action, without reaching results though (e.g., demonstration on how one uses a knife to cut a tomato, but without actually cutting it).

Note: Pantomimes and demonstrations when providing an equivalent message to what is being uttered at the same time are figurative, metaphoric in nature.

## Annex III: Metonymic Patterns

***Metonymy*** **=** Figurative equivalence between two entities that *come from the same domain, they have a similar array of associations – there is no transfer of qualities from one referent to another.*

For example, one may refer to the notion of "monarchy" and show an image of a crown, or a presenter saying "I'm in Athens" and the video showing the Acropolis on the background. These are a number of subtypes of metonymic patterns some of them well known in linguistic literature too (language metonymies).

We have a compilation of such patterns, as found in Language-Vision association naturalistic files. In the examples given below, language may express one referent and image may show another. We present these metonymic patterns into groups that reveal fundamental, conceptual relations, i.e., relations between two concepts, regardless the modality used to express them. The patterns are expressed below so that they follow an expected 'image to language' direction, however, there might be cases that the opposite direction is served by a specific pattern too. These relations are mostly action-centric (see the Minimalist Grammar of Action for a deeper understanding of all such relations that involve actions in Pastra and Aloimonos 2012). We do not claim to list all possible metonymic patterns; the granularity at which one may identify such patterns can vary substantially. What we claim though is that this level of granularity of expressing conceptual relations leads to a finite and highly economic set of relations that expresses basic, pragmatic relations between (concrete and abstract) concepts, regardless the representation modality.

## Metonymic Pattern Compilation and Clustering

### Part for Whole

- *meat – animal (e.g., pork - pig)*
- *object-component* (e.g., wing – airplane)
- *member–collection* (e.g., woman – crowd)
- *portion–mass* (e.g., piece of bread – loaf)
- *place–area* (e.g., building – city)

### Container for Content

- *container for content* (e.g., bottle – milk)

### Tool for Action

- *object/instrument/substance for action employed for* (e.g., knife – cutting)
- *entity for purpose/use (*e.g., camel – transportation*)*

### Agent for Action

- *Agent – action (e.g., butcher - slaughering)*
- *manufacturer - characteristic action* (e.g., butcher – cutting)

### Object for Action

This pattern refers to objects affected by an action.

Note: in language, there is a case of metonymies called "logical metonymy". For example, the word "book" in "I enjoyed the book" stands as a direct complement of the verb "enjoyed", instead of the omitted verb "reading", i.e., "I enjoyed reading the book". There is a substantial literature on this phenomenon, which is common in language. In our view, in such cases, an action is omitted, while its complements (tool/agent and affected object) are present, along with –in this example- a qualitative characterisation of

the action (i.e., the verb "enjoyed"). In that sense, logical metonymy is more of an 'affected object for action' metonymy. For the COSMOROE annotation, such linguistic cases have to be solved first and it is only then that one should look into the relation between language and images.

### Entity for Feature

- *Thing – defining property* (e.g., ball - round); note also that in Language we have **lexicalized feature analogies** (e.g. "ball-shaped", "heart-like"), **nominalised adjectives** e.g., "the brownish", "the blue", "the poor" and **adjectival/feature verbs**: e.g., "make square". When such phenomena are present in a multimodal context, one needs to solve the language metonymy first and then draw a multimedia relation with what is depicted e.g., "people" is the implied entity for the verbally expressed feature "poor"; then, the image of the "people" shown in the video stands in a *token-type* relation with the implied entity.

### Entity for Material

- *thing made of the material – material* (e.g., golden ring - gold)

### Entity for Measurement Unit

- *object/substance* – measurement unit (e.g., beer - pint)

### State of Entity for Entity

- prestate of thing for thing (e.g., dough - bread)

### Result for Action

- *Result - action* (e.g., pizza – pizza making, smashed potato – smashing potato)

### Trigger Action for Action

- *action that triggers – action triggered* (e.g., play music - listen); these are cases of temporal inclusion (i.e., there is some overlap between the two events and at no time can normally one event occur without the other). These are usually perception events and this specific metonymy pattern captures this fact, without resorting to temporal definitions, just to pragmatic ones.

### Action for Goal

- *action for purpose* (e.g., walks – visits); it is only language that reveals goal.

### Action for Cause

- *action for cause* (e.g., kiss - gratitude)

### Effect for Cause

- *effect for cause* (e.g., slipping – slipping ground)

### Location for Entity

- *Place for people* (e.g., island - inhabitant), it covers all cases of: inhabitants, visitors of a place/location, customers, personnel of a company/shop etc.)

### Location for Event

- *place for event (*e.g., church - wedding ceremony)

### Step for Event

- *step of a process for process* (e.g., pay – shopping, open luggage – unpack)

### Result for Event

- *result for event* (e.g., pizza – pizza making)

### Aspect for Abstract Entity

- *place for institution* (e.g., church – religion)

- *people for institution (ex: institution for people involved* (e.g., student - education) It covers only cases of institutions (θεσμοί) and the people involved in these institutions, i.e., professionals (λειτουργοί) and people served (λειτουργούμενοι, επωφελούμενοι) e.g., schools, churches etc.)

- object/artifact for art (e.g., building – architecture)

### Aspect for Abstract Feature

- *aspect for concept* (e.g., houses – poverty, running - fatigue)

*Pure metonymy vs. synecdoche*

The following are considered to be Synecdoche cases, by some people: part for whole, aspect for concept, defining property for category defined by the property, species for genus, material for thing made of the material.

Boundaries between synecdoche and metonymy are not clear and widely accepted; for the purposes of the CMR analysis, we consider all these cases metonymic.