Journal of Pragmatics 175 (2021) 112–128

Contents lists available at ScienceDirect

Journal of Pragmatics

journal homepage: www.elsevier.com/locate/pragma

“Explanation videos unravelled: Breaking the waves”

John A. Bateman*, Leandra Thiele, Hande Akin

Faculty of Linguistics and Literary Sciences, University of Bremen, Bremen, Germany

ARTICLE INFO

Article history:
Received 24 January 2020
Received in revised form 19 October 2020
Accepted 12 December 2020
Available online 3 February 2021

Keywords:
Explanation videos
Multimodality
Legitimation Code Theory
Knowledge-building
Multimodal cohesion
Empirical methods

ABSTRACT

Explanation videos are increasingly common on media websites such as YouTube and are used by school students, university students, and members of the general public alike. Such videos cover all areas of knowledge and aim to provide viewer-appropriate explanations concerning a large variety of topics. It is, however, still far from clear how such videos work and under what conditions they are effective. In this paper, we consider how this can be measured and whether guidelines can be determined empirically for their improvement. Building on the notion of *semantic waves* developed in Legitimation Code Theory, we discuss the design of cumulative knowledge-building processes and how to isolate cases where this fails to operate, selecting examples from a number of videos of this kind. To support this, we introduce a detailed annotation framework that fully reflects the multimodally-rich extent of our data. This framework systematically defines coding categories for use with the ELAN annotation tool and uses these for the adjacent construction of multimodal cohesive chain diagrams. We first motivate and describe the application of this annotation framework and then show through subsequent multimodal cohesion analyses how constructed cohesive chains can be interpreted within the scope of pedagogically relevant semantic wave patterns.

1. Introduction and research challenge

Explanation videos as we define them here are generally short, pre-recorded audiovisual presentations intended to communicate factual or procedural knowledge concerning some selected topic. Such videos show considerable diversity, ranging over non-professional productions through to specifically designed course-related content for higher education. They are generally distributed via the web, but can also be made available as part of committed online course content within institutional contexts. Although precise statistics are difficult to acquire and in any case vary across both settings and regions (cf. Bärtl, 2018; Saurabh and Gautam, 2019), educational videos already constitute a significant source of factual knowledge in many learning contexts, including various school levels and tertiary education, as well as serving the public at large. The broad growth observable in video use is due to several reasons, ranging over the increased accessibility of digital environments for producing complex audiovisual materials, the ease of dissemination of those materials, and generally increasing demands for distance learning.

This proliferation notwithstanding, there is surprisingly little agreement in the technical literature concerning just what makes a successful explanation video and, just as important, what might prevent a video from functioning for its intended purpose of communicating knowledge and skills. Despite a long history of work investigating ‘multimedia design’ (Mayer,

* Corresponding author.
E-mail address: bateman@uni-bremen.de (J.A. Bateman).

https://doi.org/10.1016/j.pragma.2020.12.009

2005, 2009), and now similarly extended bodies of empirical work in psychology and education (e.g., Craig et al., 2002; De Koning and Tabbers, 2011; Chen and Wu, 2015; Wang and Antonenko, 2017), results concerning what influences the functionality of audiovisual presentations positively remain mixed. Moreover, extensive meta-studies report conflicting results in which even basic design features (such as showing a 'talking head', combining text and images redundantly, and many more) are in some studies found to be beneficial and in others not (e.g., Eitel and Scheiter, 2015; Richter et al., 2016). As Mayer et al. conclude: "Additional work is needed to determine the conditions under which these principles apply and the underlying learning mechanisms" (Mayer et al., 2020, 837).

Bateman and Schmidt-Borcherding (2018) propose that one explanation for this continuing lack of clarity has been that the precise *discourse* placement of mobilised expressive resources is not captured adequately. This can have serious consequences for pragmatic interpretation: what might have a positive effect for understanding and comprehension in one discourse context may well, in another discourse context, prove instead to be disruptive. While promising, such an approach shifts the entire challenge to one of achieving adequate *descriptive* coverage of the extreme multimodality typically inherent to such videos. This multimodality commonly involves complex uses of spoken language, written language, visualisations in diagrammatic and pictorial form (increasingly often dynamic), as well as metacommunication of various kinds (including spoken pointers, manual gestures, dynamic presentation resources such as zooms, slide transitions, gradual text reveals, and so on) that explicitly synchronises the diverse combinations employed. In general, it is still not known how such diverse forms of information interact when combined in single coherent discourses. Moreover, it is by no means straightforward even to decide on appropriate and revealing segmentations of the data capable of supporting empirical analysis to probe such discourses further.

In this paper, we pursue these issues on two fronts. On the one hand, we introduce a detailed multimodal annotation method developed specifically for the observed complexity of explanation videos. We see the provision of annotation schemes capable of effectively supporting empirical investigation as one of the greatest challenges facing research into the pragmatic effectiveness of communication using complex audiovisual media at this time. And, on the other hand, we illustrate the utility of such a scheme by showing its application to a particular view of educational discourse proposed by Maton and colleagues within Legitimation Code Theory (LCT: Maton, 2014, 2016). By these means we seek to bring a hitherto unexplored range of new critical evaluative techniques to bear on explanation videos in order to offer further light on how to assess such videos' communicative and educational effectiveness. From a multimodality perspective, we see effectiveness in terms of the incorporation and combination of appropriate communicative modes in complementary and cohesive ways so as to ensure cumulative knowledge-building on a discourse level. We regard this as indispensable for achieving a high learning output. Moreover, although we cannot at this stage present extended corpus-based or experimental results, we consider the provision of a detailed annotation scheme linked methodologically to pragmatic interpretations relevant for knowledge communication to be an essential precondition for any subsequent advance beyond restricted case studies and towards larger scale empirical investigations.

The paper is structured as follows. First, we briefly present LCT's view of knowledge cumulation and how such cumulation has been claimed to be effectively scaffolded by appropriate patterns of information presentation. This offers a specific target for a pragmatically-oriented analysis of the materials selected. Second, we set out the organising principles of an explicit and general multimodal annotation scheme drawing on the methodologies for pursuing such studies empirically introduced in Bateman et al. (2017). And third, we show how the tighter grasp of the data achieved by this annotation framework can be employed to extend LCT's treatment of knowledge cumulation to the multimodal case. This thus serves as an illustration of how we can effectively study explanatory videos by applying descriptive principles that render the specific complex dynamic multimodal artefacts of explanation videos more accessible to empirical investigation, while still giving full attention to their multimodal aspects.

2. Cumulative knowledge building: constructing waves

In pedagogic discourses, regardless of discipline, topic and target group, one of the most significant issues concerns how knowledge can be built and disseminated efficiently. Although this has been discussed from various perspectives for particular purposes, the focus of this paper will be placed specifically on cumulative knowledge building and how discourses can be constructed to support this. In the simplest terms, cumulative learning can be described as the development of knowledge over time. Cumulative learning theory has been coined by Gagné (1968) on the premise that intellectual skills can be facilitated by decomposing into simpler skills. He also described learning processes as knowledge transfer represented in hierarchies made up by subordinate learned skills. In work on pedagogy in the New London Group tradition (New London Group, 2000; Kalantzis et al., 2010), one critical set of such skills involves 'multiliteracies', part of which in turn is the comprehension and effective use of information expressed using multiple expressive resources – such as "written and oral language with icons and still and moving images […], music and sound effects […], facial expression and hand and arm movement" (Cloonan, 2015, 99). As it is becoming increasingly common to produce educational materials for both specialist and lay audiences that demand multiliteracy, gaining better understandings of how the integration of diverse expressive resources operate in knowledge cumulation is clearly an urgent task.

One approach to exploring the relationship between particular patterns of discourse and effective knowledge cumulation has been developed by Maton and colleagues within the practical sociological framework of Legitimation Code Theory (LCT).

LCT provides several semantic dimensions of description relevant for characterising social processes of various kinds, two of which will be relevant here: *semantic gravity* and *semantic density*. Semantic gravity (SG) is defined as the degree to which meaning relates to its context, while semantic density (SD) represents the degree of condensation of meaning within socio-cultural practices. Examining classroom discourses in History and Biology classes in a secondary school, Maton (2013) observed that teaching often involved a repeated pattern of exemplifying and ‘unpacking’ educational knowledge into context-dependent and simplified meanings. It was hypothesised that such scaffolding contributes significantly to the effectiveness of communication in those educational contexts and, furthermore, might be beneficially operationalised in terms of systematic variations in semantic gravity and semantic density. These systematic variations Maton characterises as ‘semantic waves’, arguing that they offer a path to understanding how complex structures of knowledge cumulation might operate (or fail to operate).

One purpose of discussing knowledge cumulation in terms of semantic waves is to explicitly relate recognisable changes in the strengths of SG and SD to particular pedagogic strategies, such as unpacking, repacking and providing concrete examples for concepts. This promises a reliable means of analysing classroom discourse that, first, explicitly characterises forms of expression in terms of their contributions to pedagogic strategies and, second, suggests organisations for those strategies to effectively support knowledge cumulation. Interestingly for our current goals, Maton already sees these constructs as in principle operating multimodally, admitting meanings made in terms of symbols, technical terms, concepts, phrases, expressions, gestures, and even clothing (Maton, 2013, 11), although descriptive tools for extending the account beyond the verbal are still under-developed. Maton's study reveals that semantic waves may take many forms and begin and end at different levels. The kinds of semantic scales generated from the analyses of classroom interaction also indicated differences in the semantic profiles adopted which could potentially be related to the disciplines involved in that interaction.

Semantic waves track upward and downward shifts in SG and SD as a classroom discourse unfolds. The scale representing ‘downward semantic shifts’, for example, is specifically highlighted as a move

> “from highly condensed and decontextualized ideas (SG-, SD+) towards simpler, more concrete understandings, often including examples from everyday life (SG+, SD-).” (Maton, 2013, 14)

These are cases where teachers explain some complex ideas or specific concepts to students with more simplified, non-technical language and giving ‘everyday’ examples. In a Biology class, for example, the abstract term ‘cilia’ was defined as ‘the little hairs’ with the teacher additionally using body language (waving her arms). In this sense, expressing the functions of ‘cilia’ in a limited way *strengthens* semantic gravity by relying more on the context of presentation, while ‘unpacking’ the term *weakens* semantic density by depicting it only in terms of a limited number of its meanings. This kind of process is often repeated by teachers returning to the original text and finding points to ‘unpack’. However, Maton found teachers more rarely returning to the main pedagogic discourse of the subject through ‘repacking’ simplified expressions and examples into terms and ideas. Such situations are termed ‘down escalator’ profiles and were found by Maton to be potentially problematic with regard to cumulative knowledge building.

Maton offers a similar example of a semantic wave profile from a History class. Here, the teacher gives a task which begins relatively high on the semantic scale, describing “the influence of Greek and Egyptian cultures in the Roman Empire”. This calls for a decomposition of individual terms, such as ‘influence’, before beginning to address the question itself. The teacher then moves knowledge down through several specifying examples such as ‘Greek mythology’ to imply what is meant by ‘influence’ (SG+, SD-). In contrast to the previous case, however, the teacher also weakens semantic gravity by mentioning recurrent events (trade and diplomatic visits), and consequently strengthens semantic density by ‘repacking’ these examples into the technical term ‘aesthetic trade’ (SG-, SD+).

These examples from Maton's study all take place in a classroom setting, which already allows for the mobilisation of a high number of different semiotic modes, each bringing specific affordances. In other words, a classroom setting involves many instances of multimodality at work, even though Maton in his study does not particularly take this aspect into consideration and focuses primarily on linguistic, verbal modes. The overall framework he proposes appears nonetheless equally applicable to other educational genres and settings, such as the educational videos we address here. In certain respects, we might even hypothesise that the semantic waves exhibited in explanation videos need to be expressed *more* clearly and systematically than in classroom discourse, precisely because of the lack of the feedback option for resolving comprehension difficulties that face-to-face interaction naturally provides. This may then raise stricter requirements for the sensible use of the modes in this kind of medium when applied for knowledge-building that we can profitably subject to analysis. The particular modes at a video producer's disposal are the sole means available for supporting appropriate discourse pragmatic interpretations. Therefore, we now consider whether the extent to which this is being achieved may be made visible by an analysis in terms of *multimodal* semantic waves. This requires that we are able to see just which modes are pulled up and combined with each other for building such waves – which is the goal of the next section.

For illustrative purposes, we have selected explanation videos from a range of different academic disciplines addressing the following topics: a mathematical concept (statistical covariance and correlation), a linguistic concept (the phonetic processes of articulation) and a biological concept (xylem and phloem). The reason for drawing concepts from varying academic disciplines is to support further comparison across disciplines; this aspect is not, however, addressed in the current paper. Concrete examples will be taken from identified extracts from the following four videos: Video 1, ”Pronunciation and

Phonology in the EFL Classroom - Place of Articulation Part1[1]; Video 2, ”Introduction to Articulatory Phonetics (Consonants)”[2]; Video 3, ”Transport in plants - Xylem and Phloem - GCSE Biology (9-1)”[3]; and Video 4, ‘Correlation and Covariance’, a video constructed specifically by Schmidt-Borcherding for experimental investigations of effectiveness.

3. Unravelling explanation videos: from form to discourse

As we have now explained, extending the concept of cumulative knowledge-building in education to the case of explanation videos demands that the often very high degree of multimodality mobilised in those videos be made properly accessible to analysis. From a multimodal point of view, this means undertaking a fine-grained analysis of all the sensory/communicative channels involved. The ultimate goal for such an analysis is then to find out to what extent the ‘blends’ of modes in the videos manage to achieve at least formally the type of knowledge-building illustrated by Maton (2013), as this might already be regarded as an indicator for the potential effectiveness of these videos. A clear prediction for subsequent empirical investigation is then that less successful formal deployment of multimodal means should correlate with problems in knowledge development and take-up.

The first challenge, therefore, is to define means for performing this kind of detailed multimodal analysis – that is, what is needed is a strong empirical framework that takes into account all expressive modes of communication: verbal language, written language, diagrammatic visualisation (which in themselves again involve different modes), and so on. Moreover, the precise ways in which such modes are deployed within videos – including, for example, the fact that text, diagrams or other elements may be gradually revealed, partially highlighted, related visually to previous elements, etc. – need to be captured as well since this aspect of design may also play a crucial role in how meaning is being made and so condition the opportunities that recipients are given to engage with that meaning. Semiotic modes as we use the term here are identified as having both a material dimension, anchored in the biological sensory channels with which a mode is perceived and the possibilities for meaningful manipulations that a mode's material affords, and a semiotic dimension, characterising how formally classified manipulations of material may be assigned meaning. Following Bateman et al. (2017, 113f), this view differs from often too-oversimplifying definitions of modes that relate only to perceptual senses. Consequently, for example, a diagram will typically mobilise several modes making up that diagram, such as a pictorial or schematic substrate accompanied by written language and other diagrammatic elements such as arrows or lines, and so is itself already highly multimodal.

Relating this deployment of multimodal resources back to the LCT notions of knowledge building, we then seek to characterise how, for example, the lowering of semantic density and unpacking and repacking of knowledge can be achieved by appropriate choices made across all information channels collectively (and not only within the linguistic means deployed). To support this process of interpretation, working from the diverse expressive forms employed to a discourse characterisation appropriate for discussing variations in semantic gravity and density, we build on the linguistic concept of *cohesion*. Cohesion is defined as occurring whenever the interpretation of some element in a discourse is dependent on that of another (Halliday and Hasan, 1976, 5). We consequently construct cohesion analyses directly on the basis of our annotation results so as to provide an appropriate discourse description generalising over the multimodal patterns exhibited; this discourse interpretation is then, finally, related back to distinctive profiles of semantic waves as required. In this section, we discuss each of these two stages – annotation and cohesive analysis – in detail.

3.1. Audiovisual data annotation of explanation videos

Since, as noted above, the multimodal artefacts of explanation videos offer a diverse range of overlapping layers of semiotic modes (e.g., constantly changing graphics and diagrams), it is important to annotate them according to individual modes which, in interaction with the other modes deployed, are considered most likely to contribute to the creation of meaning. The extent to which meaning is actually so constructed must, in the last resort, always be addressed as an empirical question involving recipient studies. The aim of annotation within this cyclic process is, first, to draw on the state of the art in our understanding of multimodal discourse and the means by which such discourses are constructed and, second, to enable more focused empirical exploration of the consequences of combinations of meaning-making resources by providing a solid basis for formulating empirical hypotheses. Developing an annotation scheme that is capable of addressing the kind of rich multimodality at work is itself a considerable challenge. We begin, therefore, by illustrating how we have approached this task and then set out the current state of the annotation framework and its implementation. The framework is explicitly defined to be extensible, while already covering a considerable range of complex phenomena as we shall now see.

Following existing annotation practice, the developed annotation scheme consists of multiple layers of distinct kinds of information plus explicit descriptions of relationships between those layers and between elements within layers. Moreover, given that we are working with a temporally-based medium, the levels of description employed are generally linked back to the original data by timestamps. Thus, each layer of information segments an analysed video temporally with respect to some specified facets of the video's multimodal organisation. These temporally-related aspects of the annotation are well

[1] https://www.youtube.com/watch?v=q0iSDfFFvfE&t=20s.

[2] https://www.youtube.com/watch?v=dfoRdKuPF9I&t=55s.

[3] https://www.youtube.com/watch?v=R0wjTdBK77o&t=19s.

supported by the ELAN annotation software developed at the Max-Planck Institute in Nijmegen (Wittenburg et al., 2006) and so we describe here how we are employing ELAN in the annotation process with the goal of constructing subsequently a growing corpus of annotated video data.

ELAN is most commonly used for analysing face-to-face interactional data and for this, as with most tools of its kind, basic annotation functionality is already provided in terms of temporally defined annotation tracks or *tiers*. Tiers are typically used to provide segmentation of speech, co-speech gesture, proximity, and so on. In such cases, a tier can be seen as an ordered set of annotations – consisting, for example, of the phonetic description of a speaker's utterances, or of the types of gestures made by a speaker. Tiers can be either independent, in which case a tier is linked directly to the timeline of the video, or dependent, meaning they are linked to 'parent tiers'. Thus, particular segments defined temporally with respect to the video may receive additional dependent annotations, such as a phonetic transcription, a morphemic description, and a regular gloss or spelling layer as well.

The primary components of our annotation scheme were also defined using ELAN's notion of tiers. However, the situation with a multimodally-rich object of analysis such as explanation videos is rather different to the more typical annotation target of face-to-face interaction and so certain extensions needed to be made in how the resources provided by ELAN were used. We see this as reflecting some quite general concerns with annotating data of this kind and so our proposed solutions may find application well beyond the medium of explanation videos also. The principal challenge raised is the highly structured *visual* nature of the materials being annotated. This is a direct consequence of the medium of explanation videos drawing on semiotic modes that employ spatial layout, such as written language, diagrams, graphs, and so on. ELAN was not designed for such materials as traditionally those materials would be classified as spatially-structured, or 'nonlinear', data and not as temporally-structured, or 'linear', data (cf. Bateman 2014; Bateman et al., 2017, 155–158). Although the broad distinction between temporal and non-temporal data remains useful, media such as explanation videos, lectures, PowerPoint presentations and the like regularly transgress the distinction not only by employing time-based semiotic modes, such as spoken language, and spatially-based semiotic modes, such as diagrams, at the same time and in an explicitly synchronised fashion, but also by dynamicising the spatially-based media by incrementally introducing aspects of a diagram, graph, text, etc. The emerging areas of dynamic and interactive infographics are therefore also highly relevant here (e.g., Lowe and Schnotz, 2008; Weber, 2017); we consider our annotation framework potentially relevant for empirical studies in those areas as well.

The basic idea of our approach is then to translate spatial dependencies among visual units in the video into embedded, i.e., dependent, organisations of annotation tiers as defined within ELAN. The methodological principles driving segmentation (both visual and audio) of this kind is given by the notion of embedded *canvases* as defined in Bateman et al. (2017, 213–217). Multimodal analysis according to this methodology operates by dividing an object of analysis into a set of interrelated canvases and subcanvases, each of which may carry some particular range of semiotic modes. Thus, a video (audiovisual canvas) may contain within it a static diagram (visual canvas), which in turn may contain further subcanvases utilising written language, graphics, tables, and so on. Each canvas is then treated as requiring grouped collections of annotation tiers. Structural elements within a canvas form a group of annotation tiers and the multimodal properties of each element are captured by defining dependent tiers within that group. This technique also offers a natural way of including temporal development, since this is the basic mode of data representation that ELAN supports in any case, as well as providing an explicit representation of visual dependencies and their structural embeddings.

Distinct semiotic modes are captured by defining tiers of particular types. The distinct kinds of information covered by the annotation scheme currently are: written language, verbal (spoken) language, 'talking head' appearances, diagrams, arrows and lines (sometimes within diagrams, sometimes not), highlighted units, and forms of transitions in visual presentations (including 'panning', 'reframing', 'hard cut', 'dissolves', etc.). Each of these is represented by a tier of a specific type. For types of units with distinct *sub*-categories (such as types of arrows), specific coding categories are defined, expressed using ELAN's notion of 'controlled vocabularies'. This offers a more systematic way of characterising the semiotic options available and taken up. We compiled, for example, a controlled vocabulary for different forms of diagrammatic elements according to their function, ranging over connecting lines (or arrows), direction arrows, focusing arrows, labelling lines, and 'path arrows'. The temporal extent of any of these visual units is then indicated by the extents of any corresponding annotation segments within their respective tiers. Thus, the annotation naturally allows for capturing the manner of presentation of units – in particular, for example, dynamic diagrams or static diagrams in which parts of a diagram may 'appear' either as a whole or separately by employing animation effects.

Finally, several additional kinds of information that would not be appropriate to represent as distinct dependent tiers also needed to be associated with particular units in the data. For example, including a written unit's specific stylistic devices (e.g., the use of different colours, highlighting techniques, etc.) as tiers would generally lead both to an explosion of tiers and to considerable redundancy in those cases where a unit's properties do not change whenever it is used in a video. Information of this kind is recorded using ELAN's 'comments' feature whereby free-text may be associated with arbitrary elements or intervals present in the annotation. This feature is also used to assign unique identifiers to individual tiers for subsequent data processing. We capture this additional information employing standard XML representational conventions or semi-structured text rather than using actual 'free-text'; this is done in order to ease subsequent automatic processing. For example, a highlighted written language element additionally employing colour to highlight just one portion of the written text might receive the additional comment:

```
<FRAME SHAPE="SQUARE">
  <YELLOW>Difference</YELLOW>
  between Consonants and Vowels</FRAME>
```

The tag <FRAME SHAPE> accounts for the highlighting technique chosen for the written unit as a whole, here the written unit "Difference between Consonants and Vowels", namely that of using a frame (and in this case a SQUARE shaped frame). The tag <YELLOW> then accounts for the second chosen highlighting technique for the particular written sub-unit (in this case the word "difference"), namely colour coding. In each case, the element placed between tags is the respective unit being annotated. The use of XML tags in this manner makes the framework extensible, as it is then possible to employ already standardised XML representations of typography, style and so on.

The comments section was also used to document which diagrammatic elements were connecting with each other (e.g., a connecting line or arrow often connects a written unit, mostly a label, to a specific part of a diagram). Such information is crucial for appropriate descriptions of diagrams (cf. Hiippala and Orekhova, 2018) but is not naturally supported in ELAN, which only works in a time-aligned manner where spatial dimensions (such as describing parts of a specific place of a diagram) of a mode cannot be properly annotated and thus could not otherwise be incorporated into the annotation framework. Developing reliable annotation tools that more naturally support annotation of both spatial and temporal information would clearly be of considerable benefit. Providing information in a structured form within the ELAN comments field is, however, seen as a useable workaround at this time, making subsequent data comparison more feasible and traceable for follow-up research questions and more corpus-based analysis.

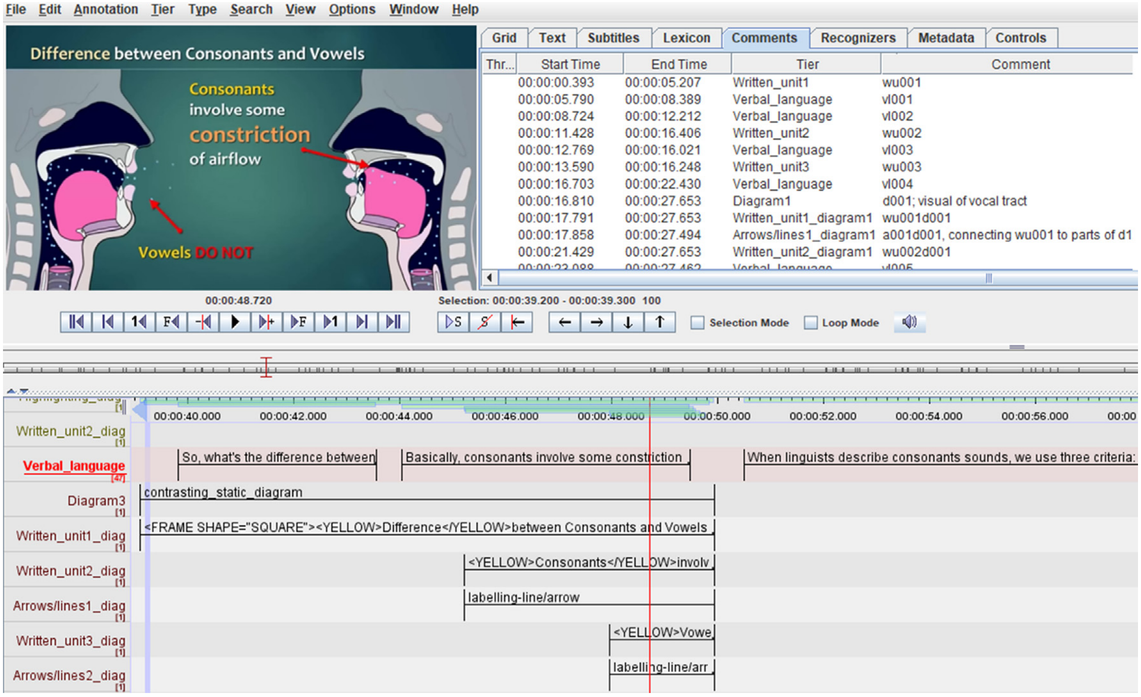
In order to clarify this annotation method, we now provide a detailed example, referring to the screenshot shown in Fig. 1. This screenshot shows annotation in progress for one of our introductory explanation videos from linguistics focusing on phonetics and phonology. As usual for ELAN, the frame upper-left shows the video being analysed, the horizontally organised rows in the lower-half of the image show a selection of the respective tiers used in the annotation, while the table shown in the frame upper-right shows the additional information specified as comments as just described.

The particular annotation in progress is from timestamp 00:00:39 in Video 2. At this point, the video introduces two visuals of a vocal tract placed facing each other (i.e., mirrored). At the same time, in the upper part of the screen the textual heading "Difference between Consonants and Vowels" appears while a voice-over states: "So, what's the difference between consonants and vowels, you might ask". This means that there are already at least three different semiotic modes at play: written language (including typography and highlighting), diagrams, and spoken language. A few seconds later, additional units appear on the screen, namely another written unit "Consonants involve some constriction of airflow" alongside a red arrow, pointing to a specific part in the vocal tract visual on the right-hand side; this thus develops further the diagrammatic nature of the depicted vocal tract. Immediately following this, a written unit "Vowels DO NOT" appears accompanied by another red arrow, this time pointing to a specific part in the vocal tract visual located on the left-hand side of the screen. The voice-over continues saying "Basically, consonants involve some constriction of airflow, whereas vowels do not". We thus see a typical example of synchronised multimodal meaning construction, fully utilising the temporal and spatial affordances of the medium and distributing information, including discourse information (such as 'contrast'), across the expressive forms available.

For the purposes of annotation, all of the semiotic modes mobilised by the video within its audiovisual canvas have to be segmented and grouped appropriately into individual independent and dependent tiers. All of the introduced modes are therefore treated firstly as individual tiers for further annotations. Concentrating on the diagram that is the visualisation of the two contrasting vocal tracts, we therefore create a tier called "DiagramX", where X is a number selected to uniquely identify the diagram – in this case it is simply the third diagram used in the video. Since this diagram is itself a major canvas, covering the entire frame and utilising further subcanvases, it is considered an independent tier with its own temporal anchoring to the timeline of the video. Independent tiers such as diagrams, verbal language and written language units that do not rely on embedding canvases are then managed within ELAN by selecting the corresponding tier type, or status, 'None', indicating no temporal embedding within a larger segment.

As described above, tiers within ELAN may also be associated with controlled vocabularies for specifying which particular subcategory of a given annotation type applies. For tiers annotating diagrams, our annotation scheme defines a controlled vocabulary made up of the categories: static diagrams, dynamic diagrams, contrasting static diagrams, and contrasting dynamic diagrams. This list may be extended as further types of diagrams are found in the data analysed; it would also be useful here to make explicit links to standardised catalogues of diagram types available in the literature (cf., e.g., Bertin, 1983; Engelhardt and Richards, 2018). For the diagram discussed here, we select the category *contrasting static diagram* as the visual part of the diagram clearly shows two different, but contrasting states of a vocal tract.

Moving on to the use of the second mode to be annotated in our example, i.e, the written text in the upper half of the screen saying "Difference between Consonants and Vowels", this is also captured by introducing a tier corresponding to the element. This tier is again given a name uniquely identifying the element, in this case: "Written_unit1_diagram3" as it is giving the diagram a title, thus it is part of this specific main canvas. We reflect this relationship between the diagram as a whole and the textual element by defining this tier to be a dependent tier, selecting "Diagram3" as its parent tier. This is where the above mentioned notion of interrelated canvases and subcanvases is made explicit; in terms of ELAN's annotations, this newly defined tier is assigned the status 'Included in' (i.e., annotation on a parent tier can be temporally divided further with gaps allowed), capturing its temporal dependence on the presence of the diagram as a whole. Since



Thr...	Start Time	End Time	Tier	Comment
	00:00:00.393	00:00:05.207	Written_unit1	wu001
	00:00:05.790	00:00:08.389	Verbal_language	vl001
	00:00:08.724	00:00:12.212	Verbal_language	vl002
	00:00:11.428	00:00:16.406	Written_unit2	wu002
	00:00:12.769	00:00:16.021	Verbal_language	vl003
	00:00:13.590	00:00:16.248	Written_unit3	wu003
	00:00:16.703	00:00:22.430	Verbal_language	vl004
	00:00:16.810	00:00:27.653	Diagram1	d001; visual of vocal tract
	00:00:17.791	00:00:27.653	Written_unit1_diagram1	wu001d001
	00:00:17.858	00:00:27.494	Arrows/lines1_diagram1	a001d001, connecting wu001 to parts of d1
	00:00:21.429	00:00:27.653	Written_unit2_diagram1	wu002d001
	00:00:23.088	00:00:27.462	Verbal_language	vl005

Written_unit2_diag [1]

Verbal_language [47]
So, what's the difference between
Basically, consonants involve some constriction
When linguists describe consonants sounds, we use three criteria:

Diagram3 [1]
contrasting_static_diagram

Written_unit1_diag [1]
<FRAME SHAPE="SQUARE"><YELLOW>Difference</YELLOW>between Consonants and Vowels

Written_unit2_diag [1]
<YELLOW>Consonants</YELLOW>involv

Arrows/lines1_diag [1]
labelling-line/arrow

Written_unit3_diag [1]
<YELLOW>Vowe

Arrows/lines2_diag [1]
labelling-line/arr

Fig. 1. ELAN interface in annotation mode illustrating the use of multiple tiers and comments for the visual representation of a phonetics video (Video 2).

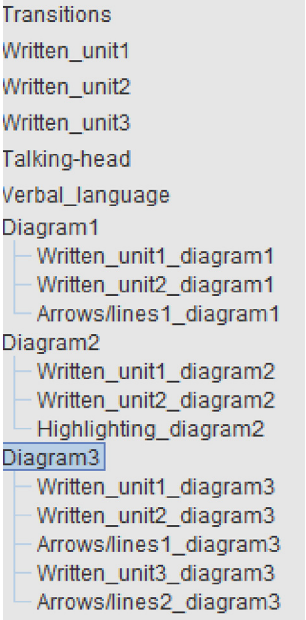
textual elements can exhibit considerable variation, we do not specify particular controlled vocabularies for these but rely instead on the comments feature that ELAN provides as illustrated above. In this way we are able to give two (or more) different kinds of information concerning the same tier: the actual orthographic representation of the written unit as well as any highlighting techniques employed. We proceed in the same way for the other two written units, introducing dependent tiers with unique labels (“Written_unit2_diagram3” and “Written_unit3_diagram3”) and marking the units' presence in the visual field of the video by appropriately delimited temporal intervals within each tier; this can be seen in the lower portion of Fig. 1.

For the two red arrows, a similar procedure is followed, first naming the tiers appropriately to provide unique identifiers, e.g., “Arrows/lines1_diagram3”. For tiers corresponding to arrows and lines, it again makes sense to employ a controlled vocabulary to capture distinct subcategories of units; these refer to the arrow's function and form as discussed earlier. In the present case, as these two arrows function to associate the written units 2 and 3 in the diagram to specific parts of the visual of the vocal tract, we correspondingly chose the category ‘labelling-line/arrow’ from the respective controlled vocabulary. Finally, in order to give the necessary further information of what exactly these arrows are referring to, additional information concerning connectivity is given in ELAN's comments field in a manner similar to that described above for typography. An example of this can be seen in Fig. 1 in the comments entry for ‘Arrows/lines1_Diagram1” (timestamp: 00:00:17.86–00:00:29.49) two up from the bottom of the comments window shown upper right: the comments field gives an identifier for the particular segment at issue and the semi-structured text description “connecting wu0001 to parts of d1”. The comments fields are also used to provide informal glosses, or labels, for the units being described in order to facilitate the readability of the subsequent annotation. Thus, the first arrow of diagram 3 receives the specification: “a001:d003, connecting wu002:d003 to part of d003 (right-hand visual tract)”, where ‘a001:d003’, ‘wu002:d003’ and ‘d003’ are shorthand identifiers for the particular temporal extents of the respective units ‘Arrows/lines1_diagram3’, ‘Written_unit2_diagram3’ and ‘Diagram3’ involved in the visualisation at that point and the phrase in parentheses offers an informal gloss of the particular part of Diagram 3 intended.

There is, however, still further information to be expressed in the current example concerning the arrows, namely (at least) their colour. In order to avoid cluttering the annotation in ELAN further and since we have already exhausted ELAN's possibilities of assigning information to a tier (the annotation of the tier itself and the use of the comments), we outsource this kind of information to a separate spreadsheet, associating properties (e.g., colour) with identifiers (e.g., a001:d003). At present this solution is used in order to proceed practically with annotation; as further experience is gained with the kinds of variability found in the data analysed, it may be beneficial to progressively migrate information of this kind to the structured comments sections. There may also be standard design choices that do not vary across the extent of a given video – such as, for example, always using red arrows. The use of an external spreadsheet as a form of default ‘style sheet’ allows us to capture generalisations of this kind as well.

Finally, the verbal language is also included in the annotation within its own tier. The tier type used for the spoken language is ‘independent’. For our current purposes, the transcription of the spoken language was kept quite simple, maintaining standard orthography and dividing segments according to audible pauses made by the speaker. As usual with annotation of interactional data, it would be possible to have multiple verbal language tiers for multiple speakers. Also, nothing excludes the inclusion of phonetically more accurate annotations, but this was not pursued as this stage.

Thus, summarising our approach, we segment the data according to the modes in which they were presented in a time-aligned manner. This partitions the audiovisual data into individual potential meaning-making units, all of which receive unique labels. Tiers are created for each unit as was shown above in Fig. 1. These tiers are grouped when particular modes ‘belong’ to the same (sub)canvas creating a hierarchy of dependent tiers correlating with the structural visual dependencies of the units in the video at any point. The hierarchy corresponding to our example in this section is shown in Fig. 2. Here we can see the independent written units followed by grouped dependent sets of tiers for three diagrams, including ‘Diagram3’ discussed above. Each diagram contains several subunits (indicated by indentation), including text elements, arrows and highlighting strategies. As explained above, each of these may develop in its own ways over time, although broader temporal inclusion relations – such as the parts of diagram 3 necessarily remaining within diagram 3 – are captured by the tier dependencies. Videos annotated in this way then provide us with units in modes including verbal language, written units, diagrams, as well as a rich variety of diagrammatic elements, such as arrows and lines, etc.



- Transitions
- Written_unit1
- Written_unit2
- Written_unit3
- Talking-head
- Verbal_language
- Diagram1
  - Written_unit1_diagram1
  - Written_unit2_diagram1
  - Arrows/lines1_diagram1
- Diagram2
  - Written_unit1_diagram2
  - Written_unit2_diagram2
  - Highlighting_diagram2
- Diagram3
  - Written_unit1_diagram3
  - Written_unit2_diagram3
  - Arrows/lines1_diagram3
  - Written_unit3_diagram3
  - Arrows/lines2_diagram3

Fig. 2. Tier dependencies defined using ELAN.

3.2. Audiovisual cohesive patterns

We next employ a multimodal cohesion analysis to investigate how the patterns of multimodal resource usage identified by our annotation can be shown to give rise to various semantic wave patterns. Revealing such processes of meaning-making formally – i.e., within any artefact analysed – is undertaken in order to support empirical explorations of how people make sense of the objects of analysis. To perform multimodal cohesion analysis, we drew from the audiovisually extended framework developed by Tseng (2013), which involves constructing a cohesive chain diagram for audiovisual data such as that we are analysing here and not, as would be the case with traditional cohesion analysis, just for verbal texts. The above described systematic fine-grained segmentation and annotation of the video data is therefore a necessary step for doing any further empirical work concerned with detecting significant patterns of meaning-making processes, including the construction of cohesion analyses.

As described above, cohesion analysis applies whenever elements of a text require interpretations of other elements of that text in order to receive their own interpretation. Re-occurrences can be either prospective, as in cataphora, or retrospective, as in anaphora. In the multimodal case, this generalises beyond linguistic re-occurrence so as to include re-occurrences in any semiotic mode. As Tseng (2013) explains, such re-occurrence relations can be of various kinds;

for present purposes we will restrict attention primarily to co-reference, synonymy, substitution and repetition. What then binds such occurrences together is the common notion of discourse, within which discourse referents may be progressively developed by suitably designed contributions making use of *any* forms of expression available to the medium.

The standard form of representation used for multimodal cohesion analyses consists of cohesive chain diagrams which visualise the exhibited re-occurrence relations in a multimodal text. Such diagrams highlight the overlapping of semiotic modes in a structured way, allowing the various contributions of multimodal resources to be tracked exhaustively across the development of a text. We argue below that a cohesion transcription of this kind provides a clearer visualisation of knowledge-building patterns and structures and so can be used directly to address our current research concerns. Without performing such a functional description, it is unlikely that the full range of interconnections functioning in a multimodal text could be made accessible to analysis. Discussions of objects of analysis lacking this level of detail have tended as a consequence to be illustrative or restricted to small sets of specific cases.

A cohesive chain diagram is organised both vertically and horizontally: items between which cohesive relations apply are set out vertically connected by lines representing the cohesive relations, or *ties*, that hold. Cohesive chains thus appear running vertically downwards. ‘Simultaneously’ present items playing roles in distinct cohesive chains are aligned horizontally. A single diagram thus typically consists of a collection of vertically-drawn cohesive chains aligned horizontally to capture co-occurrence and vertically to capture co-reference or semantic repetition. In addition, following Tseng (2013), specific types of connections *between* chains may be represented by distinguished horizontal lines. Connections drawn in verbal modes, for example, are typically the result of cohesively distinct but grammatically combined elements co-occurring in single clauses; in multimodal artefacts, such horizontal connections can also be constructed by gesture, typographical highlighting and synchronisation of spoken and written texts and diagram elements. For our analysis here, therefore, the items included within a cohesive chain diagram for a video are precisely those elements distinguished in the annotation described above. The two levels of representation – formal units and cohesive chains – are linked straightforwardly by means of the unique identifiers attributed to elements as specified in the annotation.

Although in principle any elements may be used to organise the diagram, in the particular case of the medium of explanation videos it is appropriate to take the spoken verbal language units as the main axis of development. Spoken units are strongly anchored in time and generally provide orientations for engaging with all of the other material presented. Thus, in our layout of cohesive chains here, we place numbered verbal language units on the left of each diagram, further identifying each line with selected keywords or phrases which summarise the message conveyed in that specific verbal language unit; this selection is not limited to any particular type of syntactic element. The final cohesive chain diagrams are then constructed in two stages:

(a) First, we include words or phrases which are repeated or referred to, for example, by using synonyms or substitutions, at least once during the video. Whenever a repetition or substitution occurs in immediately adjacent lines, the words or phrases are shown connected by a straight unbroken line. In cases where re-occurring expressions appear at a later stage, we use dashes to avoid interference from other chains. This then shows situations where the cohesive chains are ‘broken’ or ‘discontinuous’; such cases typically require a communicative situation to provide stronger support, such as fuller nominal phrases, particular gestures or intonation, or visual support via continuing onscreen depictions.

(b) Second, we add any temporally overlapping visual representations, such as written units or diagrams. The position of these visual units vertically among the cohesive chains indicates the time interval in which they appear. Since our primary focus is to include all modes of communication in our annotation framework, we apply the same modes in the visualisation of cohesive relations as well. In the simplest terms, this application demonstrates which units are conveyed only verbally, which only visually, and which both verbally and visually. To emphasise such modal overlaps more clearly in our graphics, units which are not only mentioned verbally, but also represented visually in written unit form, are shown underlined.

To illustrate the construction stages explained above in detail, we show in Fig. 3 the cohesive chain visualisation that corresponds to the annotations of the phonetics video that we discussed in Section 3.1 above and showed visually in Fig. 1; this picks up the video just prior to the first utterance shown there with the lecturer saying “So, just a note: we will be focusing on the phonetics of spoken languages and more specifically, consonant sounds in North American English in this video”, designated as verbal language unit ‘vl006’ in the figure and abbreviated in the diagram as “spoken languages, more specifically consonant sounds” top-left in the left-most cohesive chain. The straight vertical line above this unit shows that vl006 is already a continuation of a cohesive chain started beforehand. This same chain is then continued by the speaker verbally in unit vl007 developing two further subcategories in parallel, namely ‘consonants’ and ‘vowels’, referring back to the previous phrase. While the instructor explains the difference between these two units, they are repeated, leading us to the final part of the chain (no lines follow these after vl008).

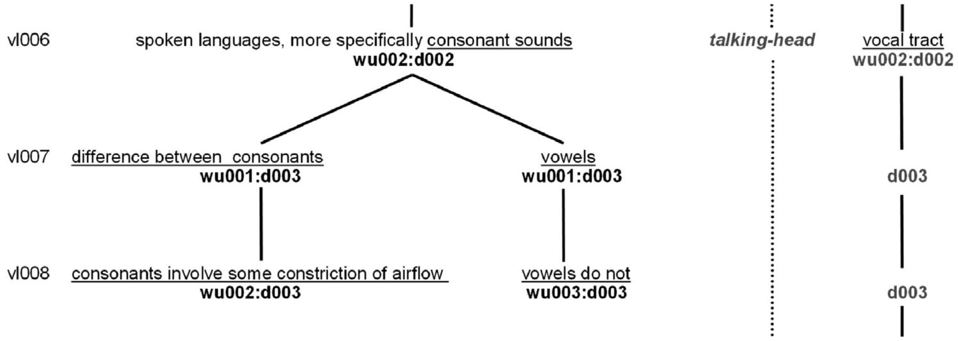

vl006
spoken languages, more specifically consonant sounds
wu002:d002
talking-head
vocal tract
wu002:d002
vl007
difference between consonants
wu001:d003
vowels
wu001:d003
d003
vl008
consonants involve some constriction of airflow
wu002:d003
vowels do not
wu003:d003
d003

Fig. 3. An extract from the cohesive chain of the annotation from Video 2 illustrated above. Here and in subsequent cohesive chain diagrams, degrees of grey shading are used to graphically illustrate the contributions of differing semiotic modes.

At the same time, the video is developing visually. The spoken utterance vl006 is produced synchronously with the second written unit of a previous diagram 2. This written unit, designated in the figure by ‘wu002:d002’, states: “How consonant sounds in North American English are produced in vocal tract [sic]”. Three more written units then make their appearance overlapping with the verbal language units vl007 and vl008. These written units are wu001:d003 (“Difference between consonants and vowels”) functioning as the title for diagram 3, wu002:d003 (“Consonants involve some constriction of airflow”) and wu003:d003 (“Vowels do not”). We can explicitly compare how these units are shown in the cohesive chain diagram with their presence in the annotation snapshot of Fig. 1 above, where they appeared as separate tiers in the lower half of that figure. Moreover, since these units echo information given in their respective co-occurring spoken units, we show them in the cohesive chain diagram underlined as explained above.

A further visual aspect of the developing video is apparent in the cohesive chain running down the right-hand side of Fig. 3. This chain tracks the presence of the mirrored diagram of the vocal tract (diagram 3, labelled d003). To begin, a discourse referent for ‘vocal tract’ is produced visually by its inclusion as a phrase in the previous written unit wu002:d002; this is then replaced by the diagram itself. Thus, although ‘vocal tract’ is not mentioned verbally in either vl007 or vl008, it is nevertheless *visually* represented in d003 in the form of the contrasting static diagram showing two vocal tracts. The repetition of the label d003 along this chain in the figure shows how it remains visually present throughout all three spoken language units vl006–vl008 and so is available for cohesive cross-chain links and pragmatic inferences drawing on those links. This is therefore a very good example of a primarily spoken cohesive chain also being maintained and strengthened by continued visual co-presence.

By these means we make it completely explicit how meaning is cumulated multimodally as the ‘text’ unfolds. The resulting visualisations and their underlying cohesive chains show particularly clearly how parts of a multimodal text are connected. As we explain in the section following, this can then be used directly to reveal changes in the semantic density of content as well as suggesting points of potential comprehension difficulty. For example, one might pose the question as to what might occur for the viewer had the cohesive chain of the ‘vocal tract’ *not* been maintained beyond vl006. Such questions can scarcely be raised without doing the transcription necessary to bring out such points of potential weakness or tension.

Fig. 4 shows the continuation of the cohesive chains depicted in Fig. 3. Following the explanation about how to differentiate consonants and vowels, the instructor introduces the various ways of describing consonant sounds by categorising them in three distinct ways. For this purpose, a title ‘describing consonants’ at vl009 is followed in the cohesive chain by three parallel branches, also carried by the written units wu002:d004 (concerning voicing), wu003:d004 (concerning place of articulation), and wu004:d004 (concerning manner of articulation). The branch for voicing is then developed further as a clause defining “voicing” in terms of “what the vocal folds are doing” at vl011. This relationship is shown in the cohesive chain diagram as the horizontal line labelled as a grammatical ‘relational process’.

The cohesive chain on the right of the figure shows the development of the supporting diagram as before. Although this diagram maintains the same visual representation of the vocal tract labelled above as d003 as its canvas, it also undergoes several major developments, such as a change in intended focus (signalled by the use of arrows), their positions and, most saliently, a change in the number of visuals the diagram includes. These developing states of the diagram are labelled in the cohesive chain figure as d004, d005 and d006 respectively. More specifically, whereas d004 is a continuation of d003, d005 and d006 are moving diagrams occurring sequentially. These dynamic developments of the original diagram express contrast by showing blinking areas on the vibrating vocal folds. Even within this single diagram, therefore, we have again a variety of modalities being deployed.

We can now return to the development of the voicing cohesive chain. At vl012, ‘voiceless’ is explained in terms of ‘open vocal folds’. This is now linked directly via the arrow al001:d005 to the corresponding state of the vocal tract diagram d005, thereby providing visual support for the spoken and written descriptions given. At vl013, this is taken further with a contrastive cohesive relationship (‘open’ *vs.* ‘vibrating’) co-occurring with “voiced sounds”, as well as the further arrow al001:d006 linking to the state of the visual diagram labelled as d006. This then also expresses the contrastive relationship visually since we see the two different articulation processes depicted one after the other together with their connected written units.

Finally, there is one further aspect of the video's organisation that we have not discussed so far. In both the previous figure and the current cohesive chain diagram, we have a cohesive chain labelled ‘talking-head’. This stands for the cases where the instructor becomes visible and contributes to the visualisation of the subject being explained. While the instructor was

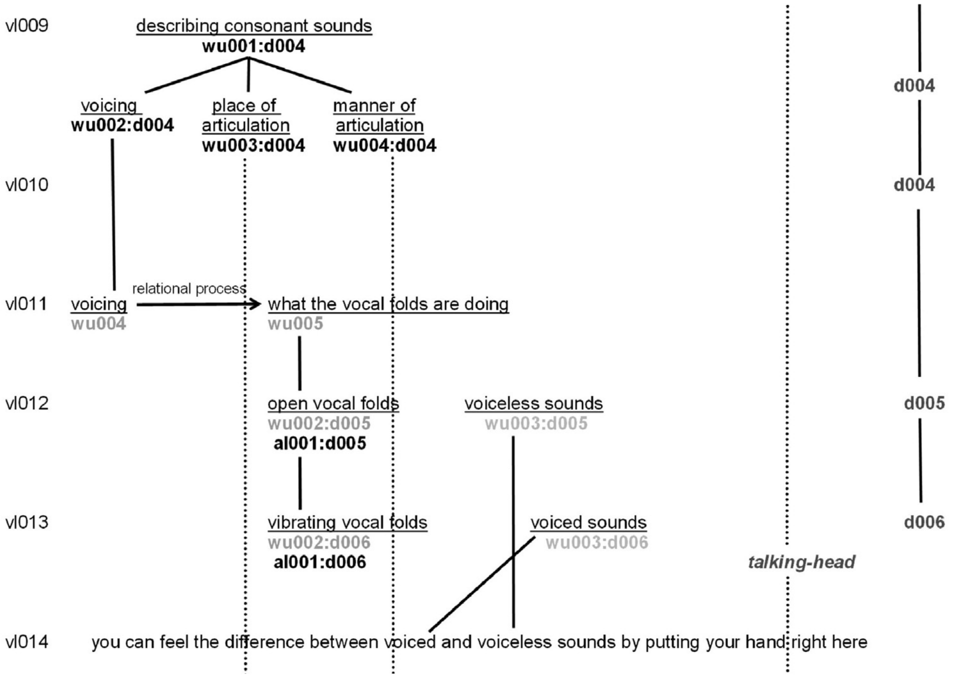


Fig. 4. An extract illustrating the synchronised combination of several modes in cohesive chains from Video 2.

present at the beginning of the video to introduce the topic, subsequently extensive use of written units and diagrams generally replaces the instructor in the visual field. This chain was included in Fig. 3 above, however, because the instructor in fact appeared during vl006 starting the sentence "So just a note..." to announce the video's focus more precisely. The next time the instructor is present is after vl013 in the current fragment and as shown in Fig. 4. Here he demonstrates how to feel the difference between voiced and voiceless sounds by placing his hand on his throat (vl014). In other words, at this point he substitutes for the diagrams identifying the location of the vocal tract – a further strikingly multimodal combination realised by cohesive cross-links across verbal language, deixis, posture and gesture.

4. Application of the annotation framework

In this section we give two kinds of examples where the kinds of detailed multimodal cohesion chains illustrated in the previous section support the direct investigation of pragmatic consequences following from an explanation video's design. The first is concerned with places of compromised 'textual' development, which we would hypothesise to be predictive of potential comprehension difficulties; we describe this case quite briefly because it is relatively straightforward but also seems quite common, although more extensive studies will be required to quantify this impression more precisely. The second kind of example addresses our more central concern, the realisation of strategies of knowledge building as introduced in Section 2 above.

4.1. Identifying places of weakness in multimodal development

We begin with one of the most straightforward cases where we might predict that a textual disfluency, here understood as ranging across all the semiotic modes used in a video, may well lead to comprehension problems. Fig. 5, for example, shows a case from our annotation of Video 3 in which a relatively technical concept is mentioned once but is not returned to subsequently.

As explained above, the diagram shows verbal spoken units whose contributions to cohesive chains is indicated by the phrases given; they are identified by the respective labels running down the left-hand side. Here we can see that this section of the video is concerned with discussing different kinds of organisms and how those organisms process oxygen. To begin, we have the spoken unit (e.g., vl001) 'what kind of organisms' and the co-occurring written unit wu001 giving a general title: "2h - Transport - Plants – Transport systems". The general category of 'organisms' is then subcategorised further as is common in technical discourse (cf. Martin, 1998); this is shown in the diagram with the horizontal relationship of a grammatical 'relational process' (i.e., a clause of identification or attribution) in vl002 relating organisms with the more specific 'unicellular' organisms. These are subsequently noted visually in written language as well (wu002). In addition, a visual image (v002) also

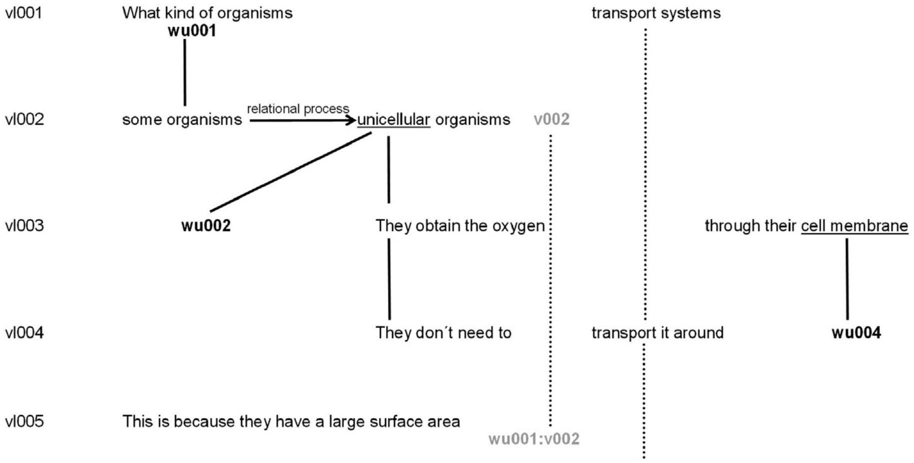


Fig. 5. An extract showing a lack of cohesive chains from Video 3. Here and in subsequent figures, talking heads are omitted for the sake of clarity when they are not interacting with other cohesive chains and so not contributing to the messages being conveyed.

appears temporally overlapping with the spoken utterance vl002. This image remains visible when wu002 (“unicellular”) is shown and so there is at least a suggestion that what is depicted in the visual might be an example of a unicellular organism.

While this component of the discourse is relatively well structured and highly cohesive, in vl003 the term ‘cell membrane’ appears, also expressed visually with the written unit wu004. However, no further explanation or examples are provided and no explicit connection is drawn to anything that can be seen at that point. Instead, after vl005, a written unit wu001:v002 (“This organism has a large surface area to volume ratio”) is added below the visual, apparently as a label. The possibility of a lack of background knowledge and the complexity of the concept ‘cell membrane’ is therefore neglected. It would have been possible to anchor this term diagrammatically into some of the visuals being employed, but this option was not taken up. The lack of cohesive connectivity here is then one clear candidate for both experimental investigation, to see if and where this lack of information leads to comprehension difficulties, and experimental *manipulation* by re-design. Achieving a broader set of annotated videos would then also offer a basis for identifying potential targets for redesign more effectively.

4.2. Showing cumulative knowledge building through cohesion

In this final subsection we show how we extend the scope of the suggested semantic wave concept to the multimodal case drawing on our multimodal cohesion analyses and the annotations these build on. To illustrate this we select some distinctive cohesive chain patterns from our explanation videos and examine the patterns more closely, relating these explicitly to the

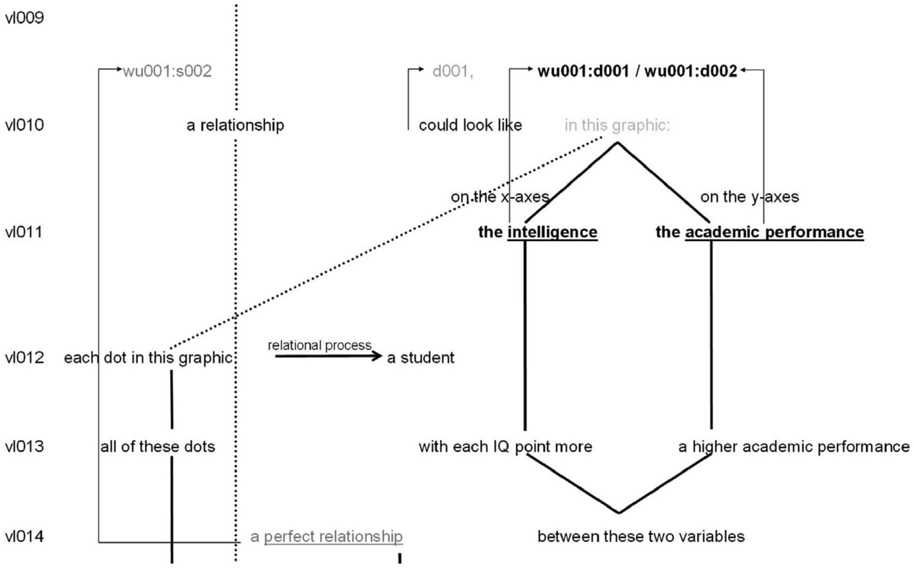


Fig. 6. An extract from the cohesive chains from a video on statistics (Video 4).

pedagogic strategies discussed by Maton (2013). This therefore seeks to unravel the knowledge building process in a clearer, and above all reproducible, fashion. Some examples from these initial findings may then offer patterns to explore on a larger scale in subsequent studies.

Consider, as our first example, the pattern of cohesive connectivity shown in Fig. 6 taken from our video introducing the statistical notion of correlations. The video extract of concern here shows an instructor explaining various aspects of the statistical procedures being discussed, at the same time showing visually a mathematical graph. Between vl010 and vl014 we find a situation comparable with the unpacking strategy reported by Maton for classroom discourse described above, although here expressed multimodally. At this particular point in the video, the instructor is introducing the relation between two variables. This is done in a technical fashion using the mathematical concepts of ‘covariance’ and ‘correlation’, which are quite abstract. Following the chains downward in the figure then shows how these are further elaborated through use of everyday language and examples: ‘variables’ are first made more concrete at vl011 in terms of ‘intelligence’ and ‘academic performance’, and then become more concrete still – taking an example from everyday life – offering “For each IQ point more, the academic performance of a student increases accordingly”.

Following this, the instructor returns to the primary purpose of explicating the graph being shown (d001) and explains further by ‘repacking’ the knowledge, thereby weakening the semantic gravity and increasing the semantic density of the presentation. More specifically, the pedagogic discourse of the subject has been grasped once again by mentioning the core concept ‘variable’ in the utterance ‘between these two variables’ (vl014). In further stages of the video, the instructor introduces the concept ‘empirical world’ saying “The real empirical world is not like this” referring to the ‘perfect relationship’ (vl014). This is again achieved by giving an everyday example from real-world situations: “We all know people whose IQ is average, achieving high scores in academic performances with diligence and effort”. Through this new unpacking process, the instructor lowers the semantic scale by moving the knowledge from an abstract sense to concrete examples, in other words, less condensed meanings.

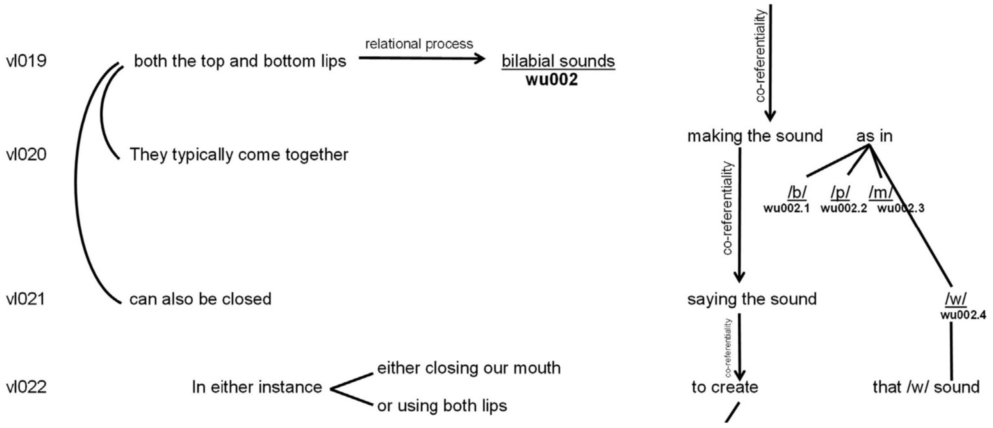


Fig. 7. An extract from Video 1 showing a further cohesive chain built on co-referentiality.

A different pedagogic strategy is illustrated in Fig. 7, in which we can see a relatively long cohesive chain at work. The video annotated here is another of our linguistics videos, again describing phonetics and phonology. In this case, the video is characterising the technical concept ‘place of articulation’. This chain, initiated at the very beginning of the video, continues throughout the segment shown in the figure. Earlier, the phrase ‘producing a sound’ was introduced and developed subsequently to fill out various kinds of linguistic events, for instance, ‘to create some sounds’, ‘making the sound’ or ‘saying the sound’, etc. The different expressions refer essentially to the same thing, namely, ‘producing a sound’, and so form a single cohesive chain. For this, we use the term ‘co-referentiality’, defined as the relationship of situational identity of reference by Hasan (1984).

Although the employment of various expressions referring to the same thing increases the possibility of cohesion in the video and therefore supports the unpacking process, it does not take an active role in repacking. However, as can be seen in vl019 and vl020 in the figure, the production of ‘bilabial sounds’ are described through everyday language: “both the top and bottom lips typically come together as we are making the sound/b/,/p/,/m/”. The replacement of the term ‘place of articulation’, which is a technical concept, with the phrase ‘making the sound’ contributes to the unpacking process. On the other hand, the instructor offers the case “can also be closed” (vl021) in which a ‘w sound’ is produced, then at vl022 it is stated that through both cases a ‘w sound’ can be created. Thus, the repacking process is partially fulfilled, because the conclusion is offered only for one sound and without returning and using the corresponding technical concept.

The previous examples show several cases of extensive content development, tracked in all cases by the appropriate unfolding of cohesive chains. However, a potential problem is raised when successive unpacking of content is *not* matched by a phase of ‘re-packing’ so that contact with the overall topic of instruction fails to be re-established – in other words, not moving back into the pedagogic discourse of the subject as discussed by Maton (2013, 13–15). This may also naturally be hypothesised as a situation where comprehension and cumulative knowledge-building may be compromised or reduced. In cases involving very complicated contexts, or when the target learner group has insufficient background information on a topic, it may be particularly important for semantic gravity to be increased and semantic density weakened. We can now also identify such situations in explanation videos directly from the cohesive chain patterns.

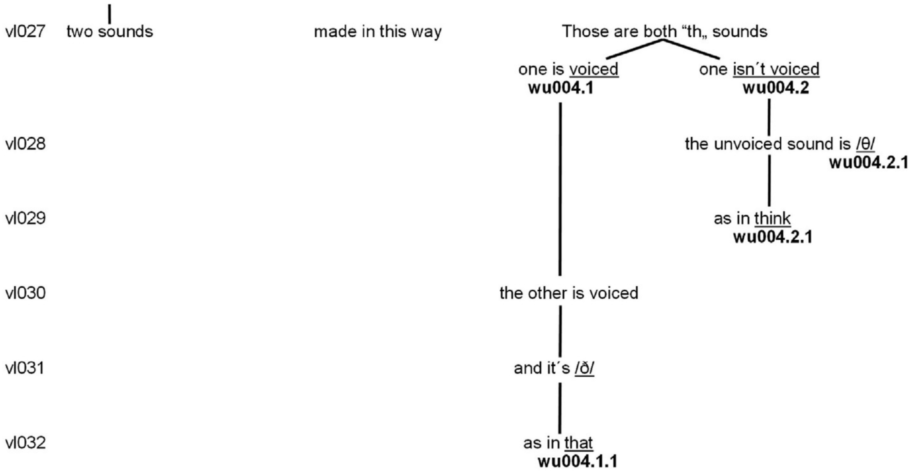


Fig. 8. An extract from the cohesive chains from Video 2 showing a case in which information fails to be ‘re-packed’.

Fig. 8, for example, shows an extract from a further linguistic explanation concerning the production of consonants. This begins with an unpacking pattern similar to those discussed above: the instructor gives one example for each concept of ‘voiced’ and ‘unvoiced sounds’. The two concepts are distinguished verbally in vl028 and two parallel cohesive chains develop them further until vl032. However, in the end, both examples remain disjoint; the instructor does not return to the start of the sequence and so no re-packing occurs. A further reason why this example is considered prominent in our analysis is the fact that this extract is the visual representation that ends the examined video. Thus, we might additionally expect the importance of summing up very briefly what has been explained up to that point, or referring back to the beginning of the core concept, to be even higher – thus suggesting the value of including broader (multimodal) genre structures for videos as a whole, which should then also be made the subject of empirical study.

Fig. 9, on the other hand, can be argued as a brief representative of both ‘unpacking’ and ‘repacking’ processes. To make this more comprehensible, we show here the spoken utterance for vl010 in its entirety. The instructor states verbally “the bit of the mouth that may get burnt when we're eating a pizza that's a bit too hot”, with intent to clarify the exact location of the ‘alveolar ridge’. This ‘everyday’ language is another attempt at strengthening semantic gravity and even might be said to productively engage embodiment and haptics. The technical term ‘alveolar ridge’, which is quite high on the scale of technicality, is ‘unpacked’ by depicting it verbally using non-technical language so that semantic density is lowered. As opposed to the lack of a ‘re-packing’ step discussed through Fig. 8, however, the instructor subsequently repeats the term ‘alveolar ridge’ within the same utterance, returning to the main subject of the discourse. In utterances vl011–vl013 the instructor then continues with the technical level development.

The patterns discussed in this section have shown how we have built a methodological framework that allows the full range of multimodal forms of expression deployed in explanation videos to be systematically related to particular patterns of cohesion, which in turn then allow description of higher level pedagogic strategies. The relative ‘completeness’ in the execution of these strategies as performing semantic waves may then serve as quite precise indicators of potential problem-spots for knowledge communication and cumulative knowledge-building. It is also made very clear why simple notions of the extent or ‘quantity’ of cohesive ties can be expected to be poor predictors of problems when considered in isolation: it is

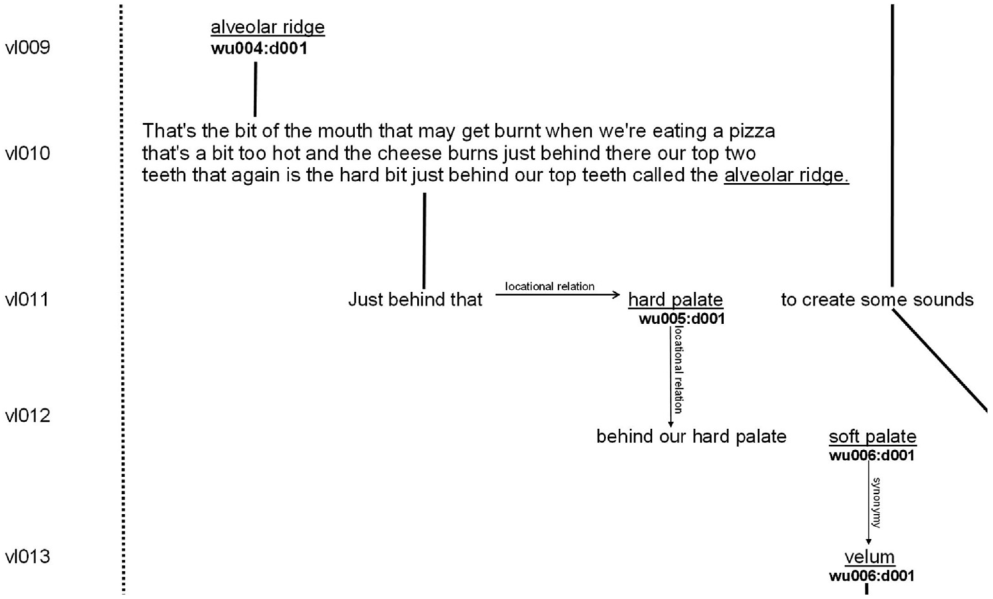

vl009
vl010
vl011
vl012
vl013
alveolar ridge
wu004:d001
That's the bit of the mouth that may get burnt when we're eating a pizza that's a bit too hot and the cheese burns just behind there our top two teeth that again is the hard bit just behind our top teeth called the alveolar ridge.
Just behind that
locational relation
hard palate
wu005:d001
to create some sounds
locational relation
behind our hard palate
soft palate
wu006:d001
synonymy
velum
wu006:d001

Fig. 9. An extract from cohesive chains exhibiting both unpacking and re-packing from Video 1.

precisely the deployment of chains in the service of particular patterns of unpacking, development, and repacking that makes the critical difference.

5. Conclusions and outlook

In this paper we have set out a methodology for investigating multimodal design factors that may have consequences for the effectiveness of audiovisual knowledge communication. In particular, we have shown how to move progressively from a fine-grained multimodal annotation scheme via patterns of cohesion to characterisations in terms of pedagogic strategies. This is intended primarily to show how expressions of knowledge may be distributed across different modalities. Collecting further empirical evidence concerning how such distributions function can also be expected to contribute substantially to multimodal research more generally, where precisely the question of how distinct modes combine is central. With complex combinations of modes it is especially important to secure analytic access to the material so as not to prejudge estimations of their effectiveness. This becomes an increasingly urgent methodological issue as the modal complexity of the objects of study grows, which is naturally encouraged by digital environments.

It is also useful at this point to very briefly contrast the approach set out here with some other current approaches to multimodality in educational contexts, including treatments of related, but distinct, media. Perhaps the most studied related medium is that of classroom interaction, as indicated in the work of Maton with which we started and analyses of lectures, both within universities and conferences, online and face-to-face (cf., e.g., Rowley-Jolivet, 2002; Shalom, 2002; Knoblauch, 2008; Karagevrekis, 2016). Although multimodality is commonly discussed, there are few principled approaches to addressing how meaning emerges from the combination of different forms. 'Modalities', considered broadly in general terms such as 'visual', 'audio', 'gesture' and so on tend to receive separate and relatively shallow analyses, with the crucial effects of combinations discussed only informally. This can be quite misleading: Karagevrekis (2016, 177–179), for example, discusses a 'mini-genre' of diagrams by attributing properties to a visual presentation which can only be made when the graphic is considered *together* with the accompanying verbal description of the lecturer.

Although this is clearly recognised by Karagevrekis, co-construction of the meaning of the presentation is not formally captured. This becomes particularly important when an audiovisual artefact under analysis exhibits design *problems* – several techniques proposed in previous approaches to multimodality, Karagevrekis' included, follow Baldry and Thibault's (2006: 48) useful proposal of segmenting according to 'phases'; however, whenever objects of analysis fail to signal phases clearly, or even provide conflicting cues, such an approach is difficult to apply. Explicitly focusing attention on just how co-construction operates and making this process accessible to empirical investigation via detailed functional annotation is then necessary, which is precisely the purpose of the account set out here. We expect, therefore, that the general approach we have outlined can find far broader application than educational videos alone.

Developing the approach further now requires the following steps: first, a representative sample of explanation videos (or related media performing a similar genre) needs to be coded using the annotation scheme described above; second, segments from the videos need to be classified according to the cohesive chain patterns described; and third, experimental evidence needs to be gathered using differences in the chain patterns as predictors of differences in teaching effectiveness. The chain patterns therefore not only offer a way of characterising discourse contexts but also isolate discourse contexts that are themselves predicted to be relevant for knowledge cumulation and so be critical for a video's success (or otherwise). In parallel research not reported here, we are also considering the use of the annotation scheme for pinpointing interdependencies between forms of expression mobilised in a video by triangulating with eye-tracking techniques; this also relates to the clear notion of discourse we employ since discourse interpretations are seen as interpretative hypotheses which should be reflected behaviourally in attention attribution (cf. Hollingworth et al., 2001; Bucher and Niemann, 2012); relations between cohesion analyses and eye-tracking are discussed in Tseng et al. (2018).

To conclude, there are many cases where we predict that the ‘textual’ (construed multimodally) building of appropriate cohesive chains will contribute to the meaning-making process, and consequently to the effectiveness of a video for knowledge communication. Cohesive chains allow us to track precisely how core concepts are developed. Examining chains constructed by co-referentiality makes visible just how that development may be distributed across modes of expression. Chains in general can then capture particular directions of semantic shifts, including downward semantic shifts from highly condensed ideas (SG-, SD+), towards more straightforward, more concrete representations (SG+, SD-), and back again. We expect examination of the distributions and consequences of such shifts to contribute substantially to our understanding of the workings of complex multimodal presentations of the kinds discussed.

Moreover, we see following from our study several practical implications for the formulation of design guidelines for educational presentations and explanation videos. When deciding on how to deploy multimodal design choices to improve learning outcomes, it is important to put any proposals for guidelines on a robust empirical basis. Such guidelines could eventually operate on both the micro and macro level and include, for example, how to incorporate, combine and arrange different modes in cohesive ways for cumulative-knowledge building so as to more effectively exploit their communicative potential to the benefit of learners' comprehension of the contents conveyed. The kinds of cohesive chain patterns that we have identified might therefore in the future provide appropriate criteria for the evaluation of videos even at the design stage. However, for this, and to avoid giving guidelines that are vague, ambiguous, or even misleading, further empirical research will be essential, particularly studies involving participant-driven experiments in order to support and refine the theoretical findings presented here.

References

Baldry, A., Thibault, P.J., 2006. Multimodal Transcription and Text Analysis: a Multimedia Toolkit and Coursebook with Associated On-Line Course. Textbooks and Surveys in Linguistics. Equinox, London and New York.

Bärtl, M., 2018. YouTube channels, uploads and views: a statistical analysis of the past 10 years. Convergence 24, 16–32. https://doi.org/10.1177/1354856517736979.

Bateman, J.A., 2014. Using multimodal corpora for empirical research. In: Jewitt, C. (Ed.), The Routledge Handbook of Multimodal Analysis, 2 ed. Routledge, London, pp. 238–252.

Bateman, J.A., Schmidt-Borcherding, F., 2018. The communicative effectiveness of education videos: towards an empirically-motivated multimodal account. Multimodal Technol. Interact. 2, 59. https://doi.org/10.3390/mti2030059.

Bateman, J.A., Wildfeuer, J., Hiippala, T., 2017. Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction. de Gruyter Mouton, Berlin.

Bertin, J., 1983. Semiology of Graphics. Diagrams, Networks, Maps. University of Wisconsin Press, Madison, WI. Translated Sémiologie graphique (1967) by William J. Berg.

Bucher, H.J., Niemann, P., 2012. Visualizing science: the reception of PowerPoint presentations. Vis. Commun. 11, 283–306.

Chen, C.M., Wu, C.H., 2015. Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. Comput. Educ. 80, 108–121.

Cloonan, A., 2015. Integrating by design: multimodality, 21st century skills and subject area knowledge. In: Cope, B., Kalantzis, M. (Eds.), A Pedagogy of Multiliteracies: Learning by Design. Palgrave Macmillan, Houndsmills, Basingstoke, pp. 97–114.

Craig, S.D., Gholson, B., Driscoll, D.M., 2002. Animated pedagogical agents in multimedia educational environments: effects of agent properties, picture features, and redundancy. J. Educ. Psychol. 94, 428–434. https://doi.org/10.1037/0022-0663.94.2.428.

De Koning, B.B., Tabbers, H.K., 2011. Facilitating understanding of movements in dynamic visualizations: an embodied perspective. Educ. Psychol. Rev. 23, 501–521. https://doi.org/10.1007/s10648-011-9173-8.

Eitel, A., Scheiter, K., 2015. Picture or text first? explaining sequence effects when learning with pictures and text. Educ. Psychol. Rev. 153–180. https://doi.org/10.1007/s10648-014-9264-4.

Engelhardt, Y., Richards, C., 2018. A framework for analyzing and designing diagrams and graphics. In: Chapman, P., Moktefi, A., Perez-Kriz, S., Bellucci, F. (Eds.), Diagrams 2018: Diagrammatic Representation and Inference. Springer, Heidelberg and Berlin, pp. 201–209.

Gagné, R., 1968. Learning hierarchies. Educ. Psychol. 6, 1–6.

Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. Longman, London.

Hasan, R., 1984. What kind of resource is language? Aust. Rev. Appl. Ling. 7, 57–85.

Hiippala, T., Orekhova, S., 2018. Enhancing the AI2 diagrams dataset using rhetorical structure theory. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan.

Hollingworth, A., Williams, C.C., Henderson, J.M., 2001. To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. Psychon. Bull. Rev. 8, 761–768.

Kalantzis, M., Cope, B., Cloonan, A., 2010. A multiliteracies perspective on the new literacies. In: Baker, E.A. (Ed.), The New Literacies: Multiple Perspectives on Research and Practice. The Guildford Press, New York, pp. 61–87.

Karagevrekis, M., 2016. Analysis of an online university lecture: multimodal perspectives. In: Gardner, S., Alsop, S. (Eds.), Systemic Functional Linguistics in the Digital Age. Equinox, Sheffield, UK and Bristol, CT, pp. 166–183.

Knoblauch, H., 2008. The performance of knowledge: pointing and knowledge in PowerPoint presentations. Cult. Sociol. 2, 75–97.

Lowe, R., Schnotz, W. (Eds.), 2008. Learning with Animation: Research Implications for Design. Cambridge University Press, Cambridge.

Martin, J., 1998. Discourses of science: recontextualisation, genesis, intertextuality and hegemony. In: Martin, J., Veel, R. (Eds.), Reading Science: Critical and Functional Perspectives on Discourses of Science. Routledge, London, pp. 3–14.

Maton, K., 2013. Making semantic waves: a key to cumulative knowledge-building. Ling. Educ. 24, 8–22.

Maton, K., 2014. Knowledge and Knowers: towards a Realist Sociology of Education. Routledge, London and New York.

Maton, K., 2016. Legitimation Code Theory: building knowledge about knowledge-building. In: Maton, K., Hood, S., Shay, S. (Eds.), Knowledge-building: Educational Studies in Legitimation Code Theory. Routledge, Abingdon and New York, pp. 1–24.

Mayer, R.E., 2005. The Cambridge Handbook of Multimedia Learning. Cambridge University Press, Cambridge.

Mayer, R.E., 2009. Cognitive theory of multimedia learning. In: Mayer, R.E. (Ed.), The Cambridge Handbook of Multimedia Learning, 2 ed. Cambridge University Press, Cambridge, MA, pp. 31–48.

Mayer, R.E., Fiorella, L., Stull, A., 2020. Five ways to increase the effectiveness of instructional video. Educ. Technol. Res. Dev. 68, 837–852.

New London Group, 2000. A pedagogy of Multiliteracies: designing social futures. In: Kalantzis, M., Cope, B. (Eds.), Multiliteracies: Literacy Learning and the Design of Social Futures. Routledge, London, pp. 9–38 chapter 1.

Richter, J., Scheiter, K., Eitel, A., 2016. Signaling text-picture relations in multimedia learning: a comprehensive meta-analysis. Educ. Res. Rev. 17, 19–36. https://doi.org/10.1016/j.edurev.2015.12.003. http://www.sciencedirect.com/science/article/pii/S1747938X15000664.

Rowley-Jolivet, E., 2002. Visual discourse in scientific conference papers: a genre-based study. Engl. Specif. Purp. 21, 19–40.

Saurabh, S., Gautam, S., 2019. Modelling and statistical analysis of YouTube's educational videos: a channel Owner's perspective. Comput. Educ. 128, 145–158. http://www.sciencedirect.com/science/article/pii/S0360131518302392.

Shalom, C., 2002. The academic conference: a forum for enacting genre knowledge. In: Ventola, E., Shalom, C., Thompson, S. (Eds.), The Language of Conferencing. Peter Lang, Frankfurt am Main, pp. 51–68.

Tseng, C., 2013. Cohesion in Film: Tracking Film Elements. Palgrave Macmillan, Basingstoke.

Tseng, C., Laubrock, J., Pflaeging, J., 2018. Character developments in comics and graphic novels: a systematic analytical scheme. In: Dunst, A., Laubrock, J., Wildfeuer, J. (Eds.), Empirical Comics Research. Digital, Multimodal, and Cognitive Methods. Routledge, London and New York, pp. 154–175.

Wang, J., Antonenko, P.D., 2017. Instructor presence in instructional video: effects on visual attention, recall, and perceived learning. Comput. Hum. Behav. 71, 79–89. https://doi.org/10.1016/j.chb.2017.01.049.

Weber, W., 2017. Interactive information graphics: a framework for classifying a visual genre. In: Black, A., Luna, P., Lund, O., Walker, S. (Eds.), Information Design: Research and Practice. Routledge, London, pp. 243–256.

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, pp. 1556–1559. http://www.lat-mpi.eu/papers/papers-2006/elan-paper-final.pdf.

John A. Bateman, *University of Bremen*. John A. Bateman is Professor of Appliable Linguistics in the English and Linguistics Departments of Bremen University. His research areas include functional linguistic approaches to multimodal document design, the semiotics of film and other media, multimodal semiotics, computational dialogue systems, formal ontology, and discourse semantics.

Leandra Thiele, *University of Bremen.* Leandra Thiele (BA, University of Bremen, 2018) is currently pursuing studies of multimodality at the University of Bremen, enrolled in the English-Speaking Cultures: Language, Text, Media degree programme. In her research, she currently focuses on Multimodality with a special interest in (Critical) Discourse Analysis in film, TV series and video games, amongst other multimodal artefacts.

Hande Akin, *University of Bremen.* Hande Akin (B.Ed., Dokuz Eylul University, Turkey, 2016) is currently pursuing studies of multimodality at the University of Bremen, enrolled in the English-Speaking Cultures: Language, Text, Media degree programme. Since she completed her bachelor's degree in English Language Teaching, her main area of research is the application of Multimodality to teaching. She is also interested in the interaction of language, image and other semiotic modes in online advertising.