# SYMBOLIC KNOWLEDGE REPRESENTATION IN TRANSCRIPT BASED TAXONOMIES

Philip Windridge, Bernadette Sharp
*Faculty of Computing, Engineering and Technology, Staffordshire University, Beaconside, Stafford, UK*

Geoff Thompson
*School of English, University of Liverpool, Liverpool, UK*

Keywords: Knowledge Representation, taxonomy, system network transcript analysis, XML

Abstract: This paper introduces a design for the taxonomical representation of participants' instantial meaning-making, as the basis for providing a measure of ambiguity and contestation. We use hyponymy and meronymy as the basis for our taxonomies and adopt the System Network formalism as the basis for their representation. We achieve an integration of transcript and taxonomy using an XML based 'satellite' system of data storage. Content data forms a 'Root' document which can then 'mapped' to by an arbitrary number of 'Descriptor' documents. This system represents instantial meanings by mapping Descriptor document elements to elements in the Root. Part of this mapping also includes the sequence of Root elements, accommodating the diachronic representation of meaning-making. This diachronic representation provides the basis for measuring ambiguity and contestation.

## 1 INTRODUCTION

This research is carried out as part of the Tracker project (Rayson *et al.* 2003) which has the aim of reducing rework through decision management. The approach described in this paper works with transcripts to analyse the active negotiation of meanings. This negotiation can lead to varying degrees of contestation and/or ambiguity between participants in the social activity from which the transcript is produced. Making this explicit, through a comparison of participants' meaning-making, has the potential to augment summative records associated with a decision, such as sets of minutes. The purpose of this paper is to introduce Transcript Based Taxonomies as a novel representation formalism.

We take a participant's experiential meaning-making as the production and construal of semantic associations between lexical units that are attributable to that participant. This aspect of meaning-making is revealed in the transcribed utterances of participants. We identify semantic associations forming dynamic taxonomical relations, both within and between these utterances, as one way of reflecting participants' negotiation of meanings. Contestation and/or ambiguity is shown by comparing participants' taxonomies where their meaning-making is comparable. The effectiveness and simplicity of this comparison is facilitated through the integration of transcript and taxonomy representations using Extensible Markup Language (XML) (W3C 2004). Transcript Based Taxonomies are the result of this integration.

In section 2 we provide a conceptual view of Transcript Based Taxonomies. In section 3 we consider the form that this representation takes by briefly discussing network based symbolic knowledge representation. Section 4 describes the representation of Transcript Based Taxonomies using XML as an integration of transcript and taxonomy. Finally, in the conclusion, we discuss this representation in terms of our research requirements and look to their future development.

The worked examples in this paper are extracted from the transcript of one formal meeting involving eight academics in total. Discussion in the meeting emphasised how project tasks would be organised and co-ordinated and included a general discussion about resource management.



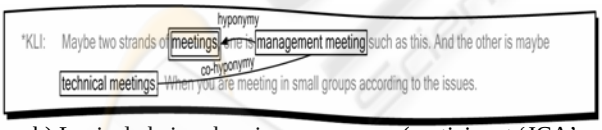
Windridge P., Sharp B. and Thompson G. (2005).
SYMBOLIC KNOWLEDGE REPRESENTATION IN TRANSCRIPT BASED TAXONOMIES.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 267-272
DOI: 10.5220/0002521102670272

## 2 TRANSCRIPT BASED TAXONOMIES

Transcript Based Taxonomies **represent meaning-making associated with a participant** in a **social activity**. They are concerned with the analysis of transcripts for two reasons. The first is that the transcript is the most practical way of accessing both the record of what was spoken and the associated contextual information affecting the meaning of utterances. The second reason, given the practicality of the transcript, is that it is the best means by which the negotiation of meanings can be analysed.

Our analysis is **based on taxonomy relations between lexical units**. Lexical units are realised by the utterance of one or more words that express a single concept; for example 'lexical unit' or 'participant's utterance'. Relations between lexical units are part of the semantic cohesiveness of the transcript and form lexical chains that can be associated with a participant. The particular types of lexical relation that we concentrate upon are **hyponymy and meronymy**. The hyponymy relation can be understood as the so-called *is-a* relation. In Figure 1a 'management meeting' and 'technical meeting' both have an *is-a* relation to 'meetings' and are co-hyponymous to each other. The meronymy relation can be understood as the *part-of* relation. For example, in Figure 1b 'set of activities', 'deliverables' and 'milestones' have been analysed as being *part-of* 'project management' and therefore in a co-meronymous relation to each other. It must be emphasised that the meaning-making we are describing emphasises the taxonomical representation of meanings as they are construed in the transcript rather than referencing any pre-defined and commonly accepted taxonomies. The meanings that we derive from the transcript are reflected in a taxonomical representation which is based upon this transcript analysis.

a) Lexical chain showing hyponymy




b) Lexical chains showing meronymy (participant 'JCA' using black lines, 'PRA' using grey lines)

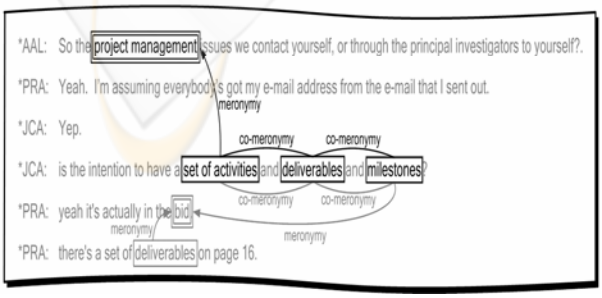



Figure 1: Lexical chains showing taxonomy relations

The lexical units used in forming Transcript Based Taxonomies are identified for their **function in experiential meaning**. Experiential meaning (Halliday 1978) can be understood as 'meaning based upon an interpretation of our experience'; in contrast to meanings that establish and maintain interpersonal relationships and meanings that have a cohesive function in a given situation. Through processes of self-reflection and communication experience is reconstituted through time so that experiential meanings change. We are apt to 'change our minds' or contradict things that we have said in the past. **Transcript Based Taxonomies must therefore be understood as dynamic**.

The participant in a social activity is an **individual person to whom a taxonomy of lexical units can be ascribed**. This ascription does not require the utterance of a lexical unit by a participant but it does require that the participant 'buy into' or accept their meaning. Figure 1b provides an example of this. The participant 'JCA' is asking about intended tasks and 'PRA' refers to existing tasks, but it can be seen that there is overlap in their respective lexical chains.

Transcript based taxonomies **allow for the measurement of ambiguity and contestation** by concentrating upon the description of relationships between lexical units. Where taxonomies can be derived for two or more participants on a given topic their comparison provides a measure of their contestation as contradictory lexical units and/or lexical unit relationships. For instance, in Figure 1b there is contestation surrounding 'operational tasks' where they have been analysed as either directly associated with 'project management' for the participant 'JCA' or with 'bid' for the participant 'PRA'. Extracting the meaning-making of a participant for a particular topic allows the analysis of the participant's meanings in terms of consistency both *in* time and *over* time. Inconsistency within a taxonomy representing temporally equivalent meanings indicates the presence of contradiction as ambiguity. For example, in Figure 1b 'PRA' appears to accept "the intention to have…deliverables" that 'JCA' talks of but then contradicts this by stating their existence in the bid as "a set of deliverables on page 16". There is a contrast between a proposal for deliverables as an *intention* and a proposal for deliverables as an *actuality*. This temporally equivalent acceptance and contradiction adds a degree of ambiguity to "deliverables" as part of "bid" on the part of 'PRA'.

# 3 REPRESENTING KNOWLEDGE

Transcript Based Taxonomies deal with symbols (as lexical units), and their interrelationships, which form a representation of participants' meaning-making. In Artificial Intelligence the development of formalisms for this purpose are part of symbolic knowledge representation. Discussing what a *representation* is, Eysenck and Keane state that "it stands for some thing in the absence of that thing" (Eysenck & Keane 1995:204). Davis *et al.* (1993) concur by saying it acts as a surrogate used for focussing on particular aspects of a thing. The *knowledge* aspect refers to the information that is contained in the representation with no claim to its veracity. As Reichgelt puts it, "the fact that some piece of information has been written down in a knowledge representation language does not by itself make it true" (Reichgelt 1991:3); the knowledge representation makes no claim about the truth of the original data. This view of knowledge representation accords with Transcript Based Taxonomies to the extent that:

- they are intended to represent that aspect of the social activity concerned with linguistic meaning-making, and
- they do not seek to make claim to any objective 'truth'.

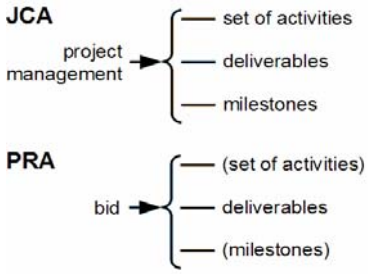
Transcript Based Taxonomies, in accordance with any other knowledge representation, afford a subjective viewpoint. This subjectivity can be seen in the inevitable ontological commitments embodied by the knowledge representation itself, in the way knowledge is organised, and in how this representation structure is populated with data when it is instantiated. The discussion of the ontological commitments embodied in Transcript Based Taxonomies, and the discussion of instantiation, is unfortunately beyond the scope of this paper and forms the basis for future publications. Here we shall simply state that Transcript Based Taxonomies are not concerned with representing 'objective' or commonly shared abstracted truth.

The taxonomical representation of lexical units lends itself to a network representation such as semantic networks. This form of symbolic knowledge representation has already been used to support natural language processing by embodying a structural linguistic approach. Simmons, for instance, considers semantic networks to be "a computational theory of superficial verbal understanding in humans" (Simmons 1973:63) which he uses for the recognition and generation of a subset of English sentence structures. Woods (1985) goes further by indicating their use as a means of understanding and modelling cognitive processes where they offer a way of "representing the meanings of sentences inside the brain (of humans or other intellects) that is not merely a direct encoding of the English word sequence" (Woods 1985:220). However, in the research reported here we are not attempting to support natural language processing, we are only concerned with representing participants' meaning-making as they occur in a transcript. This greatly simplifies the task of representation in terms of the number of relationships that are required.

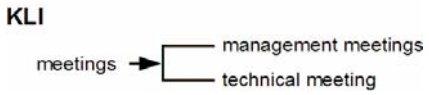
The network representation formalism adopted in this research is called the system network (Martin 1992; Eggins 1994). Figure 2 shows the system network version of the lexical chain examples taken from Figure 1. System networks share many features with semantic networks. At their most basic level they are nodes connected by arcs that can be grouped by the type of relationship they represent in a similar way to the 'and/or' graph. A system network can be comprised of one or more systems, with each system defined by one or more entry conditions and one or more outcomes. For instance, in Figure 2a 'project management' is an example of an entry condition while 'set of activities', 'deliverables' and 'milestones' are all outcomes. If 'JCA' had specified some intended 'deliverables' this would have formed a further system as part of the same system network with the outcome 'deliverables' acting as its entry condition.

A system can either represent a meronymy relation ('part-of') indicated by a brace as in Figure 2a or an hyponymy relation ('is-a') indicated by a vertical straight line connector (square bracket) as in Figure 2b. The use of parentheses, such as those surrounding 'set of activities' in Figure 2a, indicates that the node value was not directly uttered by the participant.

Unfortunately, in this basic form, the system




a) Meronymy relations




b) Hyponymy relation

Figure 2: System network taxonomy relations

network is inadequate to the requirements of Transcript Based Taxonomies because instantial meaning is lost as soon as they are abstracted from the transcript. The system network acts as a surrogate for the transcript even as the transcript acts as a surrogate for the social activity from which it is produced. While it is not possible to have direct access to a participant's meanings, because of the historically specific nature of discourse (Foucault 1972), the question is whether the transcript should be viewed as an additional layer of abstraction or whether it should be integral to the taxonomy structure. Any form of representation introduces the possibility of misconstrual. As pointed out by Davis *et al* "representations are imperfect, and any imperfection can be a source of error" (Davis *et al.* 1993:19). The burden on a taxonomy to faithfully represent meaning-making is increased if the transcript acts as an intermediary form of representation. Integration of transcript and taxonomy, thus removing a level of abstraction, will act to minimise errors.

# 4 XML REPRESENTATION

The integration of transcript and taxonomy in a single representation structure is dependent upon the way that the data is stored using XML. In the transcript it relies upon the separation of what we term the 'core data content', which is the written representation of words uttered by the participants, from data that adds meaning to or re-interprets the utterances (metadata). The 'SLA Descriptor' contains transcript specific data other than uttered words and is discussed in section 4.1. The 'Taxonomy Descriptor' contains the taxonomy relations and is discussed in section 4.2.

## 4.1 Transcripts as XML

There are two main concerns for the representation of social activities in transcripts:

1. That the representation should be as machine processable as possible,
2. That the representation should retain the 'situatedness' of meaning-making, including contextual data.

These concerns are not specific to the current research and have been addressed by a number of transcription standards. CHAT (MacWhinney 2004), for instance, is a transcription standard designed to improve the reliability and shareability of transcripts through a common, selectable and exhaustive notation. It allows the inclusion of metadata to be associated with a transcript in its entirety or to particular parts down to the lexical unit or below to the word or morpheme. This inclusion of metadata facilitates both human *and* automated linguistic analysis and processing.

The representation of CHAT transcript data in XML has been described in Clarke *et al.* (2003) and will only briefly be outlined here. The dependency of the SLA Descriptor on the Root document has already been mentioned in general terms above. This dependency takes the form of a link that 'maps' the SLA Descriptor to the Root; Figure 3 shows an example of how this is achieved. The Root document uses the `<w>` element to hold individual words and the SLA Descriptor maps onto these elements through their element number (in this case the element numbers have been added to the Root document for illustration purposes only). The co-ordinates for this mapping are identified using the attributes '`beg`' and '`len`'. The Root and SLA Descriptor documents contain enough information to provide a 'view' of the data that conforms to the CHAT transcription standard.

## 4.2 Taxonomies as XML

The XML taxonomy representation views the System Network as comprised of distinct but interconnected systems. This means that arbitrarily complex System Network structures can be represented. Figure 4 shows a simple 'single system' example first introduced in Figure 2a above. The example illustrates the scaleable notation that has been used which is based upon a simple 'entry conditions and outcomes' template for each system. Systems are represented by separate `<system>` elements under the parent element

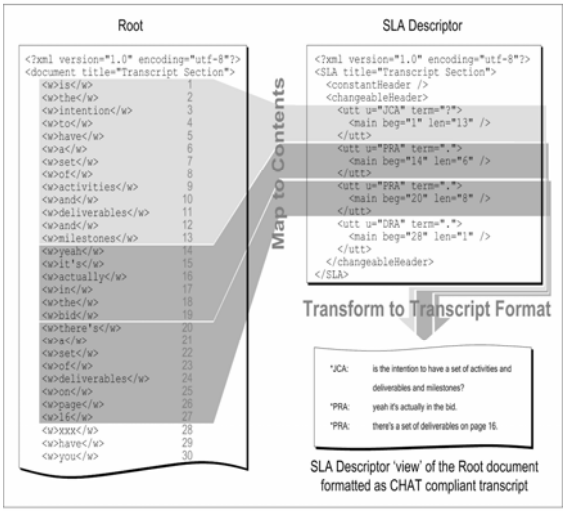



Figure 3: Reconstructing CHAT transcript from XML

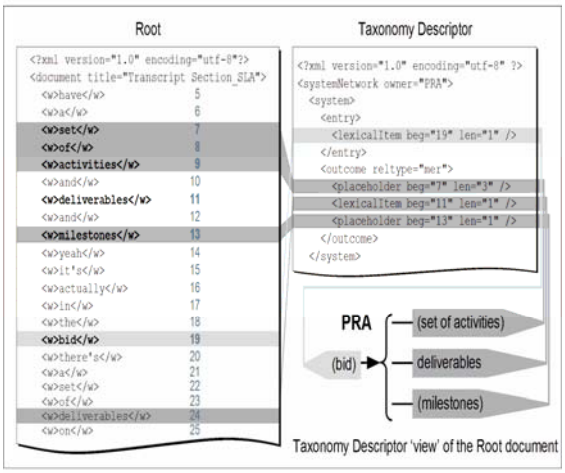



Figure 4: A System Network 'view' derived from a Taxonomy Descriptor

<systemNetwork>. Adding systems to the System Network is achieved by an outcome reference being used as an entry condition reference for a connected system. Each system contains one or more entry conditions within the <entry> element and one or more outcomes within the <outcome> element. The <outcome> element can either have an 'reltype' attribute value of 'mer' indicating meronymy or a value of 'hyp' indicating hyponymy. Both <entry> and <outcome> elements can contain <lexicalItem> or <placeholder> elements. The element can either reference a lexical unit that has been bought into by the participant or, in certain circumstances, a lexical unit that has been analysed as 'missing' from the text, for example where an object is referenced through ostension. As can be seen in Figure 4 <lexicalItem> and <placeholder> directly reference sections of the Root document using the 'beg' and 'len' attributes in the same way as described in the previous section. Using this system of pointers to record where the lexical unit originates within the transcript has two effects:

1. Sequence and proximity of lexical units are governed by their mappings to the Root document, addressing the requirement of representing diachronic meaning-making.
2. Any contextual data associated with a location in Root can be accessed in the SLA Descriptor 'view'; the same applies for the use of any other Descriptor document as a part of the representation.

However, in order for the Taxonomy Descriptor to show contestation it is necessary to extend our view of the system network. Specifically we need to show where meanings are contradictory rather than the legitimate 'mutual exclusivity' of meanings which exists in the co-hyponymy relation. There are

```
<?xml version="1.0" encoding="utf-8" ?>
<systemNetwork owner="AAL">
  <system>
    <entry>
      <lexicalItem mode="contested" beg="1937" len="2" />
    </entry>
    <outcome reltype="mer">
      <lexicalItem beg="1942" len="1" />
      <lexicalItem beg="1946" len="2" />
    </outcome>
  </system>
</systemNetwork>
```

a) Taxonomy Descriptor of participant 'AAL'

```
<?xml version="1.0" encoding="utf-8" ?>
<systemNetwork owner="JCA">
  <system>
    <entry>
      <placeholder mode="contested" beg="1937" len="2" />
    </entry>
    <outcome reltype="mer">
      <lexicalItem beg="1972" len="3" />
      <lexicalItem beg="1976" len="1" />
      <lexicalItem beg="1978" len="1" />
    </outcome>
  </system>
</systemNetwork>
```

b) Taxonomy Descriptor of participant 'JCA'

Figure 5: Contested network entry conditions

four methods that we have adopted for doing this.

Firstly, where the meaning of an entry condition is contested between two or more participants, a 'contesting' <placeholder> element references the lexical unit for all participants except the original utterer where a 'contesting' <lexicalItem> element serves this purpose. An example of this is shown in Figure 5, following Figure 1b, where the meaning of 'project management' appears to be introduced by 'AAL' referring to hierarchical lines of communication in the project, 'JCA' subsequently appears to interpret the same lexical unit as referring to organised activities of the project.

Secondly, where the meaning of an outcome is contested between two or more participants a 'contesting' <placeholder> or <lexicalItem> element is used as the outcome.

Thirdly, where the meaning of an entry condition is contested by the same participant a 'contesting' <lexicalItem> element 're-presents' the entry condition for a new system.

Lastly, where the outcomes of a system are contested by the same participant a new system is created with an identical entry condition to the original.

Effectively, the final two methods create parallel interpretations in the Taxonomy Descriptor. These interpretations are distinguished by the sequential occurrence of their lexical units in the Root document. In a similar approach, the representation of ambiguity is achieved by using an 'ambiguity' <placeholder> or <lexicalItem> element where appropriate.

# 5 CONCLUSION AND FURTHER WORK

In this paper we have discussed the representation of social activities as a taxonomy of instantial meanings which have been derived from transcript analyses. We have highlighted a number of factors that have formed the basis for this representation and have indicated their relationship to network based symbolic knowledge representation. We then outlined our XML based 'satellite system' of storing data and showed how this system is the basis for Transcript Based Taxonomies. Finally we showed how the Taxonomy Descriptor supports a number of methods for the representation of contestation and ambiguity which, together with the ability to represent the sequential development of meaning-making (see section 4.3 and the brief discussion below), provides the basis for their measure.

It has been pointed out that, due to the historically specific nature of participants' meaning-making, in addition to what can be described as subjective interpretation, it is impossible to faithfully capture intended meanings. Whilst we have accepted this limitation we have also removed an unnecessary intermediate layer of abstraction by integrating the transcript and taxonomy layers using the satellite system of XML documents. This means that the taxonomy can take full advantage of the exhaustive notation, and possibility for simplified machine processing, offered by the CHAT transcription standard. In directly mapping either an entry condition or an outcome to elements in the Root document they become identified with the instantial meanings provided by the SLA Descriptor 'view' (section 4.3). This instantial meaning is unique and any Taxonomy Descriptor that uses this mapping offers a direct and unequivocal comparison with any other Taxonomy Descriptor that maps the same point. Furthermore, this mapping carries with it a sequential order of appearance of elements in the Root document that affords a dynamic representation of meaning-making.

The primary task of Transcript Based Taxonomies is to provide a means for the comparison of meaning-making and this carries the concomitant requirement that lexical units should be associable with their synonyms, antonyms, etc., as they occur within the transcript. The association of Transcript Based Taxonomies to separate participants means that accounting for instantial synonymy, antonymy, etc., is vital for a valid comparison to take place; participants may use different words to describe the same thing, or they may use a word to directly contest another. Development of this analysis will increase the delicacy of our representation.

*This work was conducted under the auspices of the Tracker Project, UK EPSRC grant (GR/R12176/01).*

# REFERENCES

Clarke, R. J., Windridge, P. C. & Dong, D. 2003, 'Effective XML Representation for Spoken Language in Organisations', in *6th International Conference on Enterprise Information Systems*, Porto, Portugal, pp. 486-494.

Davis, R., Shrobe, H. & Szolovits, P. 1993, 'What is Knowledge Representation', In *AI Magazine*, vol. 14, American Association for Artificial Intelligence, pp. 17-33.

Eggins, S. 1994, *An Introduction to Systemic Functional Linguistics*, Continuum, London.

Eysenck, M. W. & Keane, M. T. 1995, *Cognitive psychology: a student's handbook*, 3rd edn, Lawrence Erlbaum Associates, Hove.

Foucault, M. 1972, *Archaeology of Knowledge*, Routledge, London.

Halliday, M. A. K. 1978, *Language as Social Semiotic: The Social Interpretation of Language and Meaning*, Edward Arnold, London.

MacWhinney, B 2004, *CHAT Transcription System Manual*, viewed 8 August 2004, <http://childes.psy.cmu.edu/manuals/CHAT.pdf>.

Martin, J. R. 1992, *English Text: System and Structure*, John Benjamins, Philadelphia.

Rayson, P., Sharp, B., Alderson, A., Cartmell, J., Chibelushi, C., Clarke, R., Dix, A., Onditi, V., Quek, A., Ramduny, D., Salter, A., Shah, H., Sommerville, I. & Windridge, P. 2003, 'Tracker: A Framework To Support Reducing Rework Through Decision Management', in *5th International Conference On Enterprise Information Systems*, École Supérieure d' Électronique de l' Ouest, Angers, France.

Reichgelt, H. 1991, *Knowledge Representation: An AI Perspective*, Ablex Pub. Corp., Norwood, N.J.

Simmons, R. F. 1973, 'Semantic Networks: Their Computation and Use for Understanding English Sentences', in *Computer Models of Thought and Language*, eds. Schank, R. C. & Colby, K. M., W. H. Freeman, San Francisco, pp. 63-113.

W3C 2004, *Extensible Markup Language (XML) 1.1*, Available: [http://www.w3.org/TR/2004/REC-xml11-20040204/] (4/4/2004).

Woods, W. A. 1985, 'What's in a Link: Foundations for Semantic Networks', in *Readings in knowledge representation*, eds. Brachman, R. J. & Levesque, H. J., M. Kaufmann Publishers, Los Altos, Calif., pp. 218-241.