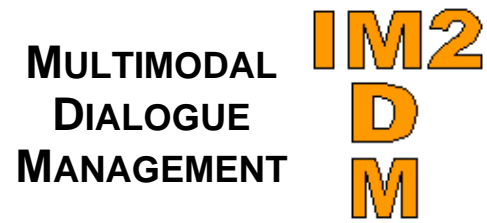




<http://www.im2.ch>



<http://www.issco.unige.ch/projects/im2/mdm/>

**ISSCO/TIM/ETI, University of Geneva**

***Approaches to Topic Annotation of Discourse in Meeting Contexts***

Eliane Luthi  
[eliane.luethi@issco.unige.ch](mailto:eliane.luethi@issco.unige.ch)

IM2.MDM Report – November 2005

# ***Approaches to Topic Annotation of Discourse in Meeting Contexts***

## ***Table of Contents***

1. Introduction
  - 1.1 Description of the data
2. Preparation of the Data for Topic Annotation
  - 2.1. Topic segmentation
    - 2.1.1 Semantic characteristics of segments
    - 2.1.2. Formal characteristics of segments
  - 2.2. Manual Keyword Annotation
3. Two Approaches to Topic Annotation
  - 3.1 The Derivational Approach
    - 3.1.1. Loss of information
    - 3.1.2. Difficulty of derivation from multiple keywords
    - 3.1.3. Impracticability if deriving from non nominal keywords
    - 3.1.4. Fluke keywords producing distorted concepts
    - 3.1.5. Need for prior disambiguation
  - 3.2. The Structure-Based Approach
4. Measuring the Need for Topic Annotation
5. An Alternative: the Keyword-Based Approach
6. Ideas for Future Work
  - 6.1. Manual annotation
  - 6.2. Automatic extraction

### ***Abstract***

This report aims to describe three approaches to topic annotation of discourse in meeting contexts: a derivational approach, a structure-based approach, and a keyword-oriented approach. It presents the advantages and disadvantages of each approach, as well as their usability for future automated annotation. It also discusses the results of a questionnaire designed to determine the need for topic annotation as it was first envisioned, and in view of these results it gives ideas for further work to be done in the domain.

## 1. Introduction

The task of topic annotation is to assign lexical labels, corresponding to topics, to previously defined thematic segments of transcribed meeting discourse. We consider topics, following Salomon (1997), as manifestations of conversational goals.

In the sphere of human-machine interaction, topic annotation can prove useful in the development of systems that analyze dialogues between humans. The exploitation of topic annotation in such systems could significantly increase the quality of user access to stored dialogue information.

### 1.1 Description of the data

The data we consider is that of discourse in meeting contexts, in particular the ISSCO meetings on furnishing a room that were staged as part of the IM2.MDM project<sup>1</sup>. For purposes of comparison, three meetings of different general topics were also studied, namely IB4010, ISSCO 22 HCI and IS1008c. In the IB4010 meeting, participants select a movie to be projected by their English movie club; ISSCO 22 discusses the acquisition of human-computer interaction resources for two separate labs; and IS1008c discusses the design of a remote control to be marketed. The data was first studied in its original format of audio and video recordings, and then in its transcribed form.

Before beginning actual topic annotation, the data was first segmented according to topics and manual keywords for each segment were then annotated. These are the preparatory steps for topic annotation, and they were performed directly on transcribed data.

## 2. Preparation of the data for topic annotation

Before we describe the preparatory steps, it should be noted that they are both highly dependent on the quality of the utterance transcription and segmentation: for topic segmentation and manual keyword annotation to be accurate and exploitable, the utterance segmentation must systematically obey the same criteria across meetings. This is especially true for work done directly on transcribed data, since utterance transcription choices induce bias into the annotator's vision of topic segments. It is therefore crucial to have coherently transcribed data. The ISSCO data was consistently transcribed according to guidelines stipulated by AMI<sup>2</sup>.

### 2.1. Topic segmentation

The goal of topic segmentation is to break up a meeting transcription into segments, which tend to reflect coherence around particular topics. Segmentation and annotation may, of course, be performed simultaneously, but it is difficult to imagine performing topic annotation on non-segmented data.

#### 2.1.1. Semantic characteristics of segments

The model chosen for our topic segmentation is hierarchical. This structure allows us to take into account broader topics as well as subtopics. This type of segmentation is in keeping with Grosz & Sidner's (1986) intentional/attentional model as well as Mann & Thompson's (1988) rhetorical structure model.

We consider that the topic segmentation must distinguish between topicalizing and non topicalizing segments. This distinction is necessary to avoid the annotation of segments that will not be sought by the user. By topicalizing we mean:

- Having the goal of pursuing a topic, and
- Being of a non metadiscursive and non metasituational nature.

Metadiscursive segments are segments containing discourse about discourse. Such segments have been labeled *topic="meta"*. An example of a metadiscursive sequence, from IS1008c (70), follows.

Ed: 'Cause I had another comment.

---

<sup>1</sup> <http://www.im2.ch> and <http://www.issco.unige.ch/projects/im2/mdm/>. IM2.MDM is funded by the Swiss National Science Foundation.

<sup>2</sup> The Augmented Multi-Party Interaction Project can be found at <http://www.amiproject.org/>

Metasituational segments are considered to be any discourse about metadata<sup>3</sup>, meeting participants or technical conditions of the meeting. We have also labeled these segments *topic="meta"*. For both metadiscursive and metasituational segments, no lower length limit was observed. All metadiscursive and metasituational utterances were subsegmented as "meta", regardless of whether they appeared as widows, as in IB4010 (73), which is composed of only one metasituational utterance:

Mirek: The picture is not very good quality.

or in conjunction with other metadiscursive and metasituational utterances, as in IB4010 (85), a metasituational segment which is interrupted and then reloaded:

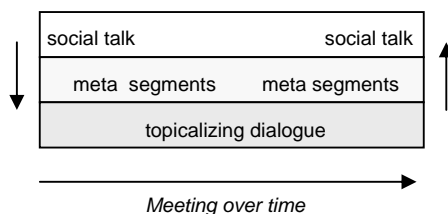
Mirek: Yeah. Could you go to the next slide?  
 [...]  
 Denis: Ah, it's actually there.  
 Agnes: Yeah.  
 Mirek: Yeah, it's here and that's actually my last slide. So um  
 Denis: Why you choose\* you choose seven movie?  
 Mirek: I don't know, because it nicely f- #! I was thinking perhaps we are not able to go through more of them.  
 Denis: Okay.

Metadiscursive and metasituational segments are not necessarily semantically homogeneous: in the above segment, the segment discusses first a slide, and then a time constraint. But because both of these sequences are metasituational, no segmentation was carried out to distinguish between them.

Lastly, the segmentation should take into account social talk, which does have linguistic content (weather, weekends, sports, etc.) but the content is irrelevant to the meeting. Social talk is characterized by two features:

- its function, which is purely social and not informative, making it a feature unique to discourse. Mazur (2004) divides conversation into behaviors related to information seeking and providing and behaviors "not specifically oriented toward information [...] neutral behaviors such as pleasantries, gossip, humorous behavior and empathic behaviors." Brown and Yule make a similar distinction when they speak of the interactional use of language, which is used "to establish and maintain social relationships (1983:1)" and is characterized by short turns. Interactional use, according to the authors, is in opposition to transactional use, which has clear topic and is characterized by longer turns.
- the semantic widening of its content from the rest of the meeting's content. Downing (2000) remarks that "the topic framework or message content is usually built up after the preliminary greetings and polite remarks have been made."

We can therefore imagine a meeting as a social interaction in three phases: talk whose function is social, metadiscursive and metasituational talk, and finally topicalizing dialogue. The process is reversed towards the end of the meeting.



<sup>3</sup> In the case of the ISSCO meetings, metadata includes the slides presented, the documents circulated, the pens used by the participants, and any other visual or textual support used during the meeting.

Social talk thus seems to be a necessary phase to the development of topicalizing dialogue, but it cannot enter the same semantic grouping as the rest of the meeting topics. We have therefore labeled these segments as *topic="unrelated"*. For example, ISSCO 35 (1) is considered social talk:

Agnes: How are you?  
Denis: Hello.  
Andrei: I'm fine.  
Agnes: How was your holiday?

### **2.1.2. Formal characteristics of segments**

It is generally agreed that discourse structure and linguistic form are mutually constraining (Passonneau & Litman 1997). Segment boundaries should coincide with topic shift and therefore should be formally marked. Formal markers that can indicate topic shift include prosodical markers such as pitch (raised at segment-initial phrases, lower at segment-final phrases), pauses of at least one second, and speech rate (accelerating at segment-final phrases [Grosz & Sidner 1986]); use of adverbial expressions such as "generally" at segment-initial phrases (Brown & Yule 1983); and use of cue phrases, including "for example", "speaking of" "but anyway" and "now back to" at segment-initial phrases, and "fine" "okay" and "that's all for..." at segment-final phrases (Grosz & Sidner 1986). If after intuitive segmentation of thematic episodes uncertainty about the exact location of a segment boundary persists, then cross-checking with formal markers is necessary. For instance, in ISSCO 35 (55-56), analysis of formal markers permitted the repositioning of the segment boundary, originally after "input", to coincide with the long pause observed after the confirming cue word "okay".

Susan: And so that was...essentially uh my input (<1 s)  
Agnes: okay (>1 s)  
Agnes: One thing that you mentioned, the black-

Despite the claim that topic shift is formally marked, Passonneau & Litman (1997) remind us that segment boundaries are fuzzy and that the use of formal markers is not systematic. This is probably due to the fact that the dynamic nature of discourse allows for in situ topic negotiating and building, which makes for difficult segmentation in comparison to text.

## **2.2. Manual Keyword Annotation**

The second preparatory for topic annotation is manual keyword annotation following basic guidelines. It should be noted that this step is not a necessity; however, manual keyword annotation gives a good semantic overview of topic segments, which is why we have chosen to perform it before proceeding to topic annotation.

For manual keyword annotation of thematic segments, we have observed three basic principles:

1. The annotation is nominal; adjectival when necessary (such as in noun-deprived segments). The need for adjectives may be a topic-specific problem, as they seem especially abundant in the ISSCO data in the form of discussions about the color, shape and design of the furniture.
2. Nominal (noun-noun) compounds may be selected when necessary. Such compounds often display semantic subordinacy to simplexes and this may a priori seem reason enough for excluding them: in Rosch's basic level theory (Rosch and Lloyd 1978), for instance, simplexes correspond to the basic level, and compounds to the level below (*kitchen table* is subordinate to *table*, *sports car* is subordinate to *car*). Similarly, Bauer (1983) analyzes compounds as hyponyms of simplexes<sup>4</sup>. However, semantically speaking, *bookshelf* – a compound turned simplex – is not more abstract than *coffee machine*, which is a compound. There is therefore no reason to exclude such nominal compounds from keyword annotation on the basis of their being too specific.

The issue to be addressed is rather the extent to which nominal compounds should be chosen over their simplex heads (*bulletin board* over *board*, *printer problem* over *problem*). We have attempted to follow a rule based on Rosch's basic level theory to analyze the need for such compounds: if the heads of these compounds are superordinate, like *area*, then it is necessary to use the whole compound as a keyword, which would then reflect basic level (*reading area*). There is some intuition to this, as well: we know that we do not reason in terms of areas and boards, and that these would be much too vague as keywords.

---

<sup>4</sup> WordNet seems to reason in this way as well. For example, magazine rack > rack; armrest > rest.

- Phrasal (adjective-noun) compounds are to be excluded. Compounds such as *red sofa* are problematic because the semantic distance they display to their nominal heads is extremely high. The exception to this rule is fixed expressions, such as *collaborative workspace*, in which one element systematically entails the other, and does not exist outside of the other. If such expressions seem semantically relevant they may then be retained for keyword annotation.

### 3. Two Approaches to Topic Annotation

On the basis of manual topic segmentation and manual keyword annotation, a methodology for accurate topic annotation was sought. The first approach explored was what we will refer to as the derivational approach.

#### 3.1. The Derivational Approach

The method used in this approach was to derive hypernyms (superordinate concepts) from keywords, either intuitively or by using a lexical resource such as WordNet.

Example (ISSCO 35 ([77]):

Denis: I suggested uh a table uh in which it's possible to store some books and magazines.  
 Denis: And this table also uh can serve as an extra seat, so it's quite nice I think.  
 Andrei: mhm

<i>table</i>	→	<i>furniture/piece of furniture</i>
keyword		superordinate concept

The advantages of this approach include its systematic methodology that allows for high inter annotator agreement as far as manual annotation goes. It also presents a high probability of computational feasibility, which is desirable for future automated annotation within the MDM project. However, the approach presents several disadvantages, including significant loss of information, the difficulty of deriving one topic from multiple keywords, its impracticability if deriving from adjectives or proper nouns, the creation of fluke keywords inducing distorted concepts, and the need for prior disambiguation of keywords.

##### 3.1.1. Loss of information

When manually deriving concepts from nominal keywords, all non nominal information is lost. For instance, the keyword *space* that appears in a segment sequence *reduce space* may produce a concept that is semantically very distant from the meaning of the sequence. Moreover, this kind of annotation risks the generalization and repetition of information-poor superordinate concepts, such as *furniture*, which is a concept that can cover many different keywords. Topic annotation using this method thus risks providing very little information about topics. For example, three adjacent segments containing different keywords run the risk of all producing the same topic label, blurring any semantic distinction between them.

Example:

<i>table</i>	→	<i>furniture</i>
<i>cabinet</i>		<i>furniture</i>
<i>bookcase</i>		<i>furniture</i>

##### 3.1.2. Difficulty of derivation from multiple keywords

Secondly, deriving from several keywords at once can be problematic, if not impossible. We have noted the reoccurrence of groups of heterogeneous keywords that have no obvious semantic or cognitive link. For instance, in ISSCO 36, segment 73<sup>5</sup> yields the following keywords: {*ugliness, drawback, budget*}. To obtain a concept from these three keywords, we would be required to select a single keyword for future derivation, thus exacerbating the problem of information loss.

<sup>5</sup> See Annex I for the full transcribed segment.

### 3.1.3. Impracticability if deriving from non nominal keywords

The third problem is that the derivational approach requires strict, noun-only keyword annotation, since it is impossible to derive concepts from adjectives or proper nouns (in the automated stage). Thus, the following, cognitively obvious link cannot be made:

adj. *red*  
keyword  $\longrightarrow$  N. *color*  
concept

The repeated need for keyword annotation of proper nouns (such as in IB4010, where names of directors and actors abound) would likewise hinder automated derivational topic annotation.

### 3.1.4. Fluke keywords producing distorted concepts

The fourth problem appears when segments that are poor in nouns or rich in anaphora produce non representative keywords. The derivation from non representative keywords can distort the topic annotation, as in IB4010 (167):

Andrei: So now let's vote.  
Andrei: Should we vote just for one film or should each of us say- give-  
Mirek: So eliminate first eliminate first? Mm-hmm.  
Denis: Or we can eliminate first one one movie each, and then we s- and then we select.

Using the derivational approach with WordNet, the lone keyword *film* would give the distorted topic *show*, whereas intuitively, the concept here is (*voting*) *procedure*.

### 3.1.5. Need for prior disambiguation

One last issue to consider is the need for prior disambiguation of keywords when using this approach. WordNet makes several semantic distinctions such as *lamp*: an artificial source of visible illumination, as opposed to *lamp*: a piece of furniture holding one or more electric light bulbs. The first would give us *device* as a concept, the second *furniture*. To avoid concept distortion, disambiguation must be performed systematically for each keyword, even recurring keywords (as *lamp* could take on the first meaning in one part of a meeting, and the second in another).

The derivational approach therefore seems tailored to the simplistic rendering of topics with automated means, but seems much less semantically accurate than is desired.

## 3.2. The Structure-Based Approach

Because of the drawbacks of the first approach, a second approach that was independent of keywords was explored, which we will refer to as the structure-based approach. This is the approach adopted for the final annotations of the ISSCO data.

The approach is based on the observation of the discursive structures of the ISSCO data, which led us to note the existence of a recurring semantic pattern which can be described as follows:

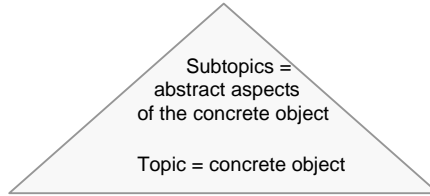
- Phase 1: statement of a topic
- Phase 2: gradual topic pursuit of different aspects of the topic

A similar frame was proposed by Grosz and Sidner (1986) when they speak of embedding relationships in linguistic structure, which the authors deem a manifestation of a deeper, non-linguistic intention or purpose. In this view, dialogues are often organized according to sets of purposes and subpurposes; a topic is thus a manifestation of a purpose, and a subtopic of a subpurpose.

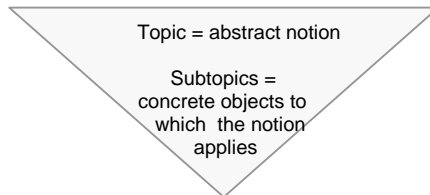
The flow of semantic abstraction between topics and subtopics can be directed upwards (towards the abstract) or downwards (towards the concrete), thus creating two basic variations of the structure. Certain concept words, such as *furniture*, invite bidirectional abstraction (towards the concrete in the form of topics such as *chair*; towards the abstract in the form of topics such as *arrangement*).

The two variants of the structure-based approach can be visualized as follows:

**Upward semantic abstraction**



**Downward semantic abstraction**



If we accept this structure as recurrent and typical, we can intuitively annotate in accordance with such a frame. For example, ISSCO 36 (44)<sup>6</sup> can be characterized as follows:

- Topic = table
- Subtopic 1 = dimension
- Subtopic 2 = purpose
- Subtopic 3 = quantity

Thus, the first subtopic, *dimension*, is to be understood as the dimension of the *table*; it is an “aspect of” *table*. Each subtopic will then be related to its main topic in an “aspect of” relationship, and each subsubtopic to its subtopic in a second “aspect of” relationship, and so on until the exhaustion of topic levels.

The main advantage of this approach is therefore the possibility of taking several levels of abstraction into account by linking topics to subtopics, making the annotation information-rich, reducing semantic distances between topic words and their corresponding segments, and allowing for a better semantic overview than the first approach. The approach is also cognitively accessible (i.e. more tailored to human cognizance) because it is keyword-independent, mirroring human deduction capacities, and the frame proposed may allow for relatively high inter annotator agreement (although this has yet to be verified).

The main disadvantage of the structure-based approach is its doubtful computational feasibility, since no link is necessarily detectable between keywords and topics. Furthermore, if we wish to explicitly state the semantic relationships between keywords and topics, complex descriptions would be needed to fully describe them (combinations of hyponymic, meronymic, and telic relationships, for instance). For example, ISSCO 36 (68) yields the keywords {*advantage, foot, table*}, but the intuitive concept annotated for the segment was *comfort*. It is possible to fully describe the lexical link between these, but the process would be time consuming and cumbersome.

It also has yet to be verified whether this structure is topic-independent or topic-specific, although annotation was successfully carried out for IB4010 and IS1008c using the approach. A topic-specific structure would of course make it difficult to define guidelines for high inter annotator agreement.

All in all, the structure-based approach appears more geared towards a human reading of annotations, but does not (at least for the moment) facilitate progress towards the future automation of topic

<sup>6</sup> See Annex I for the full transcribed segment.

annotation, mainly because of the lack of consistent and reliable means to detect links between keywords and topic words.

#### **4. Measuring the Need for Topic Annotation**

The difficulties encountered with regards to the relationship between keywords and topic words led to a renewed interest in the need for topic annotation as it was first envisioned, i.e. topic annotation that makes use of semantically abstract labels, as opposed to annotation that solely makes use of keywords.

In order to evaluate the extent to which human beings tend to describe discourse segments using topic words, two versions of a questionnaire were drawn up. Each questionnaire contained five short segments of discourse in meeting contexts, followed by a word bank. The word bank contained all nouns appearing in the segment (in singular form), as well as concept words pertaining to the segment that did not appear in the text. We sought to include equal amounts of keywords and concept words. The words were then put in alphabetical order so as to minimize priming effects.

The first questionnaire, whose segments were derived from IB4010, was filled in by 37 non English mother tongue (non EMT) older university students. The second, whose segments derived from ISSCO 36, was completed by 70 non EMT first-year university students and 6 academic staff members. No control was made for age groups, although it was assumed that the first year students' English was of a lower level than that of the older students.

Respondents were asked to read each segment and spontaneously choose up to three words from the bank that they felt accurately described the segment. They were also permitted to add in a word if they felt one was missing.

The questionnaires were collected and then analyzed in terms of use of concepts words (CW) versus use of keywords (KW). What is of most interest to us is the combinations used to describe segments, and the frequency of all-keyword or all-concept word combinations versus the frequency of combinations that contained both keywords and concept words.

Results are shown below; the questionnaires are annexed. The asterisks indicate words that are keywords; the others are to be understood as being concept words.

It should also be noted that the information gathered with the help of the questionnaires is of a general nature; the respondents were not asked to imagine themselves being users of the system the MDM project is currently developing.

## Keyword and concept word use

### Questionnaire 1: IB4010 meeting

	Box 1	Box 2	Box 3	Box 4	Box 5	All Boxes
% of available KW	50	50	60.9	37.5	55.5	53
% of chosen KW	42.9	43	65.3	41	66	51.6
Average # of chosen W per box	2.838	2.892	2.730	2.703	2.865	2.805
% of 3W combinations chosen	86.5	91.9	75.7	73	83.8	82.2
% of KW-CW-CW answers	43.8	58.8	25	40.7	25.8	39.5
% of KW-KW-CW answers	28.1	17.7	46.4	44.4	45.2	35.5
% of 3W combinations containing both KW & CW	71.9	76.5	71.4	85.2	71	75

### Questionnaire 2: ISSCO 36 meeting

	Box 1	Box 2	Box 3	Box 4	Box 5	All Boxes
% of available KW	52.9	50	50	50	46.2	50
% of chosen KW	49.6	47.6	47.7	49.8	43	47.5
Average # of chosen W per box	3	2.763	2.816	2.851	2.972	2.88
% of 3W combinations chosen	92.1	81.6	80.3	89.1	88.9	86.4
% of KW-CW-CW answers	44.3	35	36.1	37.9	42.2	39.3
% of KW-KW-CW answers	38.6	35.5	32.8	42.4	39	37.8
% of 3W combinations containing both KW & CW	82.9	71	68.9	80.3	81.3	77.1

### Recapitulative table: combined results of Questionnaires 1 and 2

	All Questionnaires
% of available KW	50.6
% of chosen KW	50
Average # of chosen Ws per box	2.86
% of 3W combinations chosen	85
% of KW-CW-CW answers among 3W combinations	39.37
% of KW-KW-CW answers among 3W combinations	37.05
% of 3W combinations containing both KW & CW	76.42
% of 3W combinations containing both KW & CW among all answers	65

## Most used words

### Questionnaire 1: IB4010 meeting

	Box 1	Box 2	Box 3	Box 4	Box 5
Most used W	film / plot	commemoration	history*	duration	poster*
2 <sup>nd</sup> most used W	crime* / story*	liberation*	movie / film	Apocalypse Now*	movie*
3 <sup>rd</sup> most used W		celebration* / concentration camp		version*	cast

### Questionnaire 2: ISSCO 36 meeting

	Box 1	Box 2	Box 3	Box 4	Box 5
Most used W	configuration*	armchair* / comfort	table*	comfort	dimension
2 <sup>nd</sup> most used W	arrangement	furniture	purpose	choice*	space*
3 <sup>rd</sup> most used W	space*		paper*	price*	table*

The results thus show that generally, keywords were used to describe segments roughly half the time (48% of the time for the furniture meeting, and 52% of the time for the movie club meeting). This is perhaps not surprising given that they also made up half the available words in the word bank.

What is more telling is that respondents overwhelmingly opted to use more than one word to describe a segment and that, in most cases (65% of the time), they used a three-word combination of keywords and concept words<sup>7</sup>. The predominant combination employed for both questionnaires was two concept words and one keyword, which was used 39% of the time with respondents using three-word combinations, whereas the combination of two keywords and one concept word was employed 37% of the time by those using three-word combinations.

This leads us to believe that human understanding of discourse topic is complex and naturally tends towards multi-word descriptions. What seems to be happening is the evaluation of words that appear in the text against a semantic importance scale, combined with words that are inferred from the text, which are often words semantically more abstract than those appearing in the segment. For instance, segment 5 of Questionnaire 2 yields “dimension” as the most used label, whereas the only nouns explicitly mentioned in the segment that are linked to this concept are “centimeter” and “space”.

It thus seems a gross reduction of human cognitive behavior to seek to label entire discourse segments with one concept word as was first envisioned.

### **5. An Alternative: the Keyword-Based Approach**

The questionnaire findings, combined with the difficulties encountered when exploring the derivational approach, have led us to consider adopting a purely keyword-based approach in view of future automated exploitation<sup>8</sup>. This approach focuses solely on keywords and eliminating concept words from the annotation altogether. This implies extracting quality keywords for each segment – minimizing fluke keywords – and finding a way to determine which keyword would most likely best represent the segment. This approach is information-rich, allowing access to several semantically representative words at once, and it also seems to be in sync with user habits (since users are used to performing keyword searches on the Internet, for example).

For example, in IB4010 (138):

Denis: Okay, I actually wanted to propose you some posters related to this movie.

Denis: Uh maybe it w- maybe it's going to open the discussion on on posters as well.

The keywords are {*poster, movie, discussion*}. A good keyword extractor would allow us to determine that of these three keywords, *poster* best reflects the topic of the segment.

### **6. Ideas for Future Work**

If we choose to adhere to the latter approach, then further work must be done on keyword annotation. Manual keyword annotation can be completed using the simple guidelines stipulated in 2.2; this annotation information can later be used to improve an automatic keyword extractor with linguistic knowledge, such that extraction is not simply an elimination of stopwords.

We can define keywords, following Hulth et al. (2001), as a “small set of terms selected to capture the content of a document.” The purpose of keyword extraction is thus to “identify (in a way: understand) the content of many forms of textual communication.” (Hunyadi, 2001).

#### **6.1. Manual annotation**

In manual keyword annotation, a typical way to reduce inter annotator variation is to first define a set of permitted keywords and then establish certain instructions for annotation. We must take care to note, however, that “the idea of limiting semantic variation to a discrete and predetermined set of well defined terms [...] is of course a dramatic simplification of human linguistic behaviour.” (Hulth et al., 2001).

---

<sup>7</sup> It should be noted, however, that the wording of the instructions (i.e. “circle up to three words”) may have incited respondents to select three words instead of one or two, especially with non EMT respondents. This is perhaps the main default of the questionnaire.

<sup>8</sup> We have not excluded the possibility of the structure-based annotations being exploitable to this end, although it is currently not clear how this might one day be possible.

Without stipulating predefined keywords, we can continue to use the previously stated guidelines. Concentrating on noun phrases seems to be a guideline which is confirmed by the literature on keyword assignment: Hulth et al. (2003) note that “when inspecting manually assigned keywords, the vast majority turn out to be nouns or noun phrases with adjectives [...] the research on term extraction focuses on noun patterns.”

Once manual annotation has been done for a significant amount of data, machine learning from manual keywords could be carried out. In this case, two main issues (following Hulth et al., 2003) are defining potential words and establishing the features of these terms that are considered discriminative. Hulth suggests the following discriminative features: term frequency, collection frequency, relative position of the first occurrence, and the part of speech (PoS) tags assigned to the term. This of course would need to be verified in view of our data pertaining to discourse and not text.

## **6.2. Automatic extraction**

A first preprocessing step could include annotating words with PoS tags and morphologically normalizing them before extraction as suggested by Hulth et al. (2001); this would help with extraction according to discriminative syntactic features.

To determine these discriminative syntactic features, we can follow Hulth’s (2003) approach on PoS tag patterns: we can, for instance, establish a list of the frequent syntactic patterns found in manual keyword assignment. Hulth found AdjNs, NsNs, AdjNpl, NsNpl, Ns to be the most frequent patterns in her experiments. This approach has the advantage of not arbitrarily restricting the number of terms that a keyword can contain. It also allows for compounds and adjectives to serve as a basis for future automatic extraction.

It is necessary to address the issue of compounds, which is a language-specific problem (concerning English and the Romance languages, as opposed to German and the Scandinavian languages). Semantically speaking, it should be possible to extract nominal compounds so as not to have information-poor, highly superordinate keywords. Words such as *area* or *board*, for instance, should not be extracted over concise basic level keyword compounds. It also seems that human annotators have a tendency to include multi-word expressions: in van der Plas et al.’s (2004) controlled manual annotation experiment, multi-word expressions were selected from a restricted list by her human annotators 26% of the time. Thus it seems to make sense to allow for multi-word expressions to be extracted, and to determine which syntactic structures should be extracted in accordance with manual annotation examples.

A second important issue is that of proper nouns. Here, adding a name-recognition module (as suggested by Hulth et al., 2001) can improve extraction since proper nouns sometimes play significant roles in certain types of text (cf. IB4010).

There is also a need for a disambiguation module that can deal with plural/singular semantic distinctions (such as the case of *nineties* vs. *ninety*).

Finally, a general semantic classification of keywords may help to establish an understanding of the way keywords relate to one another. Hunyadi (2001) suggests that associating “each lexical item with possible semantic domains together with proper morphological and syntactic parsing will assign certain relevant semantic/conceptual relations to phrases.” The use of a lexical resource such as EDR or WordNet (as was done by van der Plas et al. [2004]) could help us associate keywords to semantic domains, as well as calculate semantic distances between keywords. Of course, the danger here is recreating problems similar to the ones discussed in 3.1-3.5; once again, we should make sure to avoid associating concept words with specific segments and use this technique simply to provide general information about the semantic domains of the keywords.

However, any extraction approach that is adapted from text-based extraction techniques needs to be adjusted to take the discursive character of the data into account. The frequency of certain syntactic patterns in speech could quite possibly differ greatly from the textual frequency of these same patterns.

## References

- Bauer, L. (1983). *English Word Formation*. Cambridge University Press, Cambridge.
- Brown & Yule. 1983. *Discourse Analysis*. Cambridge University Press, Cambridge.
- Downing, Angela. 2000. "Talking Topically." In *Text and Talk*. Universidad de Castilla-La Mancha.
- Grosz & Sidner. 1986. "Attention, Intentions, and the Structure of Discourse." In: *Computational Linguistics*, 12 (3): 175-204.
- Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, pp. 216-223. Sapporo, July 2003.
- Hulth, A., Karlgren, J., Jonsson, A., Boström, H. & Asker, L. (2001). Automatic Keyword Extraction Using Domain Knowledge. In: *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2001)*. Mexico City, February 2001. LNCS 2004, Springer.
- Hunyadi, L. (2001). Keyword Extraction: aims and ways today and tomorrow. In: *Proceedings of the Keyword Project: Unlocking Content through Computational Linguistics*.
- Mann, W. & Thompson S. 1988. "Rhetorical Structure Theory: towards a functional theory of text organization." *TEXT*, 8: 243-281.
- Mazur, J. (2004). *Conversation Analysis for Educational Technologists: Theoretical and methodological issues for researching the structures, processes, and meaning of on-line talk* (2nd ed.). Lawrence Erlbaum Associates, Mahwah.
- Passonneau R. & Litman D. 1997. "Discourse Segmentation by Human and Automated Means." In: *Computational Linguistics* 23 (1): 103-139.
- Rosch, Eleanor; Lloyd, Barbara. (1978). *Cognition and Categorization*. LEA, Hillsdale.
- Salomon, P. (1997). "Conversation in Information Seeking Contexts: A Test of an Analytical Framework." In: *Library & Information Science Research* 19 (3): 217-248.
- Van der Plas L., Pallotta V., Rajman M. & Ghorbel H. (2004). Automatic Keyword Extraction from Spoken Text. A Comparison of Two Lexical Resources: EDR and WordNet. *LREC'2004 (Fourth International Conference on Language Resources and Evaluation)*, Lisbon, Portugal, vol.VI, p.2205-2208.

## ***Annex I: Meeting segments***

### *ISSCO 36 (73)*

Susan: And extremely ugly.  
Martin: yeah  
Martin: Yeah, that's - that's (unintell) ugliness.  
Susan: {laugh}  
Andrei: Well, that's the drawback, I mean it has to have a drawback {laugh}.  
Agnes: yeah  
Martin: I wouldn't even have - want to have that in my kitchen so -  
Agnes: Didn't say it was pretty but we're working on a budget.  
Andrei: yeah  
Andrei: mhm  
Agnes: Um .. so actually this whole set comes out to just under fourteen hundred.  
Andrei: Oh, great.  
Martin: Oh, uh -  
Susan: Yeah, but I think you can find nicer looking chairs than that at any discount store in town.  
Agnes: Probably, I just -  
Andrei: mhm  
Agnes: I took that one cause it was a set and it made life easier but -  
Agnes: if we want to go looking for cheaper individual chairs then -  
Martin: Ah, okay, so the difference with mine is the .. price of the armchairs.  
Agnes: why not.  
Susan: yup  
Susan: Yeah, anyway the chairs isn't what's going to break our budget uh -  
Martin: okay  
Agnes: no  
Andrei: mm  
Andrei: sorry

### *ISSCO 36 (44)*

Andrei: Okay, what about finding - sorry Martin - a smaller table.  
Martin: And at -  
Andrei: We had uh an idea last time .. it's a - a set of three side tables that you can move very easily.  
Andrei: And you can even put one on the side of the - the sofa.  
Andrei: And I think a smaller table is - is fine because we don't need such a big table.  
Martin: yeah  
Susan: Well, I - I think - I - I like the idea of a big table in front.  
Martin: I -  
Andrei: So if you don-  
Susan: It just doesn't need to be so deep.  
Susan: It could be uh - uh not uh - it could be less deep uh.  
Martin: mmm  
Martin: That's true .. but -  
Andrei: mhm  
Martin: Yeah, that - that was the other - my first other option was these uh you know these tables that can pile - you pile up -  
Susan: Little sidetables, yeah.  
Martin: on the - on the others then you -  
Martin: you - you have the different - the three different sizes.  
Andrei: mhm  
Susan: yup  
Martin: So you can use them the way you want.  
Susan: mmm  
Martin: You could have the small one on the side if you want to put just a glass on it or something like that, and the - the bigger - bigger one - the biggest one on - in - in front of the sofa.  
Susan: Well, no but the biggest one in front of the sofa is this - this dinky little table.  
Martin: yeah, that's - that's another -  
Susan: It - it's - it's no good uh -  
Martin: Yeah, that's -  
Andrei: yeah  
Martin: That's why.  
Susan: So, so one - one table -  
Andrei: Well, they are not that small.  
Susan: Oh come now, it's uh about this big.  
Martin: Yeah, it's - it's roughly seventy nine.

Susan: (unintell)  
 Andrei: Yeah, but what would you put on such a table?  
 Susan: yeah  
 Andrei: Your -  
 Susan: Well, you - you put uh - you put - you put uh - uh -  
 Andrei: The article you're reading, your legs?  
 Martin: Y- y- ya- your - your -  
 Martin: Your paper.  
 Susan: yeah  
 Martin: You're reading a paper so you have to -  
 Andrei: mhm  
 Susan: A glass of water, your cup of coffee, uh you know.  
 Andrei: mhm  
 Martin: And if you are - if you are .. if you are alone it's okay, but .. I think that .. we - we - we should have some idea about h- - on average how many people together might be in this, and I would say two, probably.  
 Martin: Because, more than two I don't think that it would be really comfortable.  
 Andrei: mhm  
 Martin: Two people, if - if one person is sitting in the armchair the second person is sitting on the sofa, and you have this tiny table -  
 Martin: It's not very - not very useful.  
 Andrei: mhm  
 Susan: yeah  
 Martin: Well, you could say, then you have the tables on the side.  
 Andrei: mhm  
 Martin: You can put things on the side.  
 Martin: Yeah, I - I - I would agree that - that's one - that's an option still.  
 Susan: mmm  
 Andrei: yeah  
 Andrei: Okay, so we need to choose a table either .. biggish or smallish.  
 Susan: yeah  
 Susan: But I still like the - I still like the stackable sidetables, because then you can move the chair out, you know you don't have to look at the people on the couch, you move the other way, and you have your little side table there to put down your .. cup of coffee or your glass or water or whatever.  
 Martin: Yeah th- I -  
 Martin: okay  
 Andrei: mhm  
 Martin: okay  
 Martin: I agree, I agree.  
 Andrei: mhm  
 Martin: That was my first option, and then I somehow -  
 Susan: Yeah, s- so you have both.  
 Martin: Yeah  
 Susan: yup  
 Martin: I agree with both.  
 Andrei: Okay, maybe we need to vote or - or whatever.  
 Martin: No, well, it's okay.  
 Andrei: So, yeah.  
 Andrei: And .. I would really argue for a .. bigger, well, I don't know a sitting table or small desk on the left and um- uh well, maybe seeing - hi Agnes .. uh, but seeing your presentation may be able to clarify, if you are for instance for or against the table on the left side so -  
 Agnes: Sorry about that.  
 Agnes: okay  
 Susan: We - we - we were just discussing this uh - the coffee table that spans most of the couch, or the sidetable .. and we're sort of in between the two.  
 Agnes: mhm  
 Susan: My proposal is that we do have a coffee table, not, perhaps as deep as what Martin shows in his picture there, or his diagram, but that we do have that and then we still have the set of the three square sidetables.  
 Agnes: right  
 Agnes: okay  
 Martin: Uh, for sure having the small tables is more flexible in a sense because you can do more things with them.  
 Susan: yeah  
 Andrei: mhm  
 Andrei: yeah  
 Martin: That's - that's clear.  
 Martin: That's uh -  
 Agnes: Okay, so -

## Annex II: Questionnaires

### Questionnaire 1: Topic Annotation of Discourse in Meeting Contexts

Please read each meeting excerpt and then circle **up to three** words that you feel **best describe** the excerpt from the list of words below. If you feel a word is missing, you may add it in the blank. Do not think it over too long, be spontaneous.

1. **Agnes:** And it's basically about five criminals who get set up. Um they're arrested, they're released from prison and the get together um and plan a crime, and basically the story is told through the views of one of the suspects. Um the tag line is kind of neat, five criminals, one line-up, no coincidence and you get told the story um as uh - that these f- characters are being controlled um by one sort of uber-villain, and in the end you find who the uber-villain was.

\_\_\_\_\_

action character coincidence conspiracy crime criminal criminality  
event film line-up movie narration plan plot prison scenario  
setup story suspect tag line uber-villain view \_\_\_\_\_

2. **Mirek:** I'm uh thinking whether there was some good occasion, uh something what has happened recently. There was something like a celebration of the fifty years after the war or something recently, no?

**Andrei:** Um I think there was a celebration of uh sixty years from the liberation of the camps, actually.

**Denis:** Sixty. S- sixty years.

**Mirek:** No, that's - sixty, yeah. S- it can't be fifty, sure.

**Agnes:** Mm-hmm.

**Andrei:** Because they were liberated beginning forty five and uh -

**Mirek:** Yeah, so.

\_\_\_\_\_

celebration camp concentration camp commemoration death camp event  
liberation occasion remembrance war World War Two year \_\_\_\_\_

3. **Andrei:** Um so my proposal which is only a proposal, would be to to keep the focus on um uh more or less history and action, so Lawrence of Arabia was, okay, about this guy during the First World War, so - it happens, okay, outside Europe and the States, but uh it's a piece of history. Uh Apocalypse Now it's uh Vietnam, if I'm um correct

**Agnes:** Mm-hmm.

**Mirek:** Mm-hmm.

**Andrei:** and, okay, Amadeus, that's much uh much earlier and it's not about - it's not violent, which is a nice thing I think.

\_\_\_\_\_

action Amadeus Apocalypse Now country Europe epoch  
film First World War focus genre guy history proposal  
Lawrence of Arabia movie narration plot setting States story line  
type Vietnam violence \_\_\_\_\_

4. **Andrei:** it's a bit long too, it's nearly three hours. But I don't know, Apocalypse Now is quite long too I think  
**Agnes:** yeah  
**Denis:** Yeah, it is. There is two – there are two versions I think, and – but w- but we – yeah, we projected the longer one, so.  
**Andrei:** I don't remember.  
**Agnes:** Yeah, the redux version is longer.  
**Andrei:** Mm-hmm.  
**Agnes:** Yeah  
**Andrei:** Yeah.

\_\_\_\_\_

Apocalypse Now duration hour length movie projection time version \_\_\_\_\_

5. **Denis:** okay, actually I wanted to propose you some posters related to this movie. Uh maybe it's going to open the discussion on on posters as well. So actually here you – on this – the first poster you can see the actors from the movie. I don't know if you know some of them,  
**Agnes:** Mm-hmm.  
**Denis:** so Jeff Bridges at the top, uh Goodman is the second one, the b- the big guy. Then Turturro, and uh uh the last one is Buscemi, and I don't remember the name of the girl.  
**Agnes:** Mm it's Julianne Moore, isn't it? Yeah.  
**Andrei:** Um-  
**Denis:** Yeah, Julianne Moore.

\_\_\_\_\_

actor actress Buscemi cast design discussion film girl Goodman  
identity image Jeff Bridges Julianne Moore movie photograph  
picture poster Turturro \_\_\_\_\_

## Questionnaire 2: Topic Annotation of Discourse in Meeting Contexts

Please read each meeting excerpt and then circle up to three words that you feel **best describe** the excerpt from the list of words below. If you feel a word is missing, you may add it in the blank. Do not think it over too long, be spontaneous.

1. **Martin:** So, uh, if you have this - the spaces is like this, and uh you have the door, and you have the - the windows. For me because of the configuration of the place .. uh, this is very difficult to use to sit because it will be a - people will be going through and if you go- if you want to go on the - on the - on the balcony, it's probably the way that you will be going.  
**Martin:** You will be coming from the door through this window out. So to me, the cozy place in this - in this room is .. this part. So -  
**Andrei:** mhm  
**Susan:** Martin can you use another colour pen?  
**Andrei:** correct  
**Susan:** I can barely see anything.  
**Martin:** Um - so the cozy part is this one.

arrangement balcony colour comfort configuration corner coziness door  
layout noise part pen place seating setup space window \_\_\_\_\_

2. **Andrei:** Uh, something important, I think it comes from your drawing, is that both Agnes and Susan wanted arms - armrests for the - the armchairs and the sofa.  
**Susan:** armchairs  
**Andrei:** And yours seem to have them uh?  
**Martin:** Yeah, that - that - I - I - yeah, yeah, yeah, sure but that - that- that- that's good.  
**Susan:** yup  
**Andrei:** mhm

armchair armrest comfort drawing furniture relaxation seating sofa \_\_\_\_\_

3. **Andrei:** Yeah, but what would you put on such a table?  
**Susan:** yeah  
**Andrei:** Your -  
**Susan:** Well, you - you put uh - you put - you put uh - uh -  
**Andrei:** The article you're reading, your legs?  
**Martin:** Y- y- ya- your - your - your paper.  
**Susan:** yeah  
**Martin:** You're reading a paper so you have to -  
**Andrei:** mhm  
**Susan:** A glass of water, your cup of coffee, uh you know.  
**Andrei:** mhm

article coffee cup drink furniture glass leg need object  
paper purpose reading support table use water \_\_\_\_\_

4. **Agnes:** Um, the armchairs – cream isn't the most ideal choice but...for this price it's...the optimal choice.  
**Andrei:** It's alright.  
**Agnes:** it's got the armrests. It's a little bit of a bucket seat but I –  
**Susan:** I – I love these seats {laugh} oh this is really comfortable {laugh}.  
**Martin:** {laugh}  
**Agnes:** I actu- I actually don't mind them cause I tend to just sort of...curl into it and not actually use the armrest as an armrest but more as a back support. But...that's just me.  
**Andrei:** mhm

armchair    armrest    back support    bucket seat    colour    comfort    choice    price  
relaxation    posture    position    seat    seating    shape    \_\_\_\_\_

5. **Susan:** And I have – I have a bit of a problem with those uh...triangular tables too.  
**Andrei:** Yeah, I don't like them.  
**Martin:** yeah  
**Susan:** N-no because...um...uh it's – it's not – well, it can be a taste matter, that's not the point, it's that um they take up essentially as much space as the square ones do and you've just lost half the space.  
**Martin:** See Susan, but still I –  
**Agnes:** But we – do we really need that much table space? Really? I mean the tables are fifty-five centimetres –  
**Martin:** Ah – but still – yeah that's right but –  
**Agnes:** right? So it's –

centimetre    design    dimension    furniture    issue    matter    problem    shape  
side table    size    space    table    taste    \_\_\_\_\_