Multimodal Corpora:
Computer vision and language processing (MMC 2016)

Workshop Programme

09:00 – 09:15 – Welcome!

09:15 – 10:30 – Oral Session 1

Emiel van Miltenburg: *Stereotyping and Bias in the Flickr30k Dataset*

István Szekrényes, Laszlo Hunyadi and Tamas Varadi: *The Multimodal HuComTech Corpus: Principles of Annotation and Discovery of Hidden Patterns of Behaviour*

Michael Amory and Olesya Kisselev: *The Annotation of Gesture Designed for Classroom Interaction*

10:30 – 11:00 Coffee break

11:00 – 12:15 – Oral Session 2

Minghao Yang, Ronald Böck, Dawei Zhang, Tingli Gao, Linlin Chao, Hao Li and Jianhua Tao: *"Do You Like a Cup of Coffee?" - The CASIA Coffee House Corpus*

Paul Hongsuck Seo and Gary Geunbae Lee: *A Corpus for a Multimodal Dialog System for Presentation Controls*

Michael Tornow, Martin Krippl, Svea Bade, Angelina Thiers, Julia Krüger, Ingo Siegert, Sebastian Handrich, Lutz Schega and Andreas Wendemuth: *Integrated Health and Fitness (iGF)-Corpus - ten-Modal Highly Synchronized Subject-Dispositional and Emotional Human Machine Interactions*

12:15 – 13:45 – Lunch break

13:45 – 15:00 – Oral Session 3

Claire Bonial, Taylor Cassidy, Susan Hill, Judith Klavans, Matthew Marge, Douglas Summers-Stay, Garrett Warnell and Clare Voss: *A Robotic Exploration Corpus for Scene Summarization and Image-Based Question Answering*

Anna Matamala and Marta Villegas: *Building an Audio Description Multilingual Multimodal Corpus: The VIW Project*

Jindřich Libovický and Pavel Pecina: *A Dataset and Evaluation Metric for Coherent Text Recognition from Scene Images*

15:00 – 16:00 – Posters and demos

Ian Wood: *Thinspiration and Anorexic Tweets*

Emer Gilmartin, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan and Nick Campbell: *Collecting a human-machine, human-human, and human-woz comparative social talk corpus*

Kristin Hagen, Janne Bondi Johannessen, Anders Nøklestad and Joel Priestley: *Search and Annotation Tools for Heritage Language Spoken Corpora*

16:00 – 16:30 Coffee break

16:30 – 17:45 – Oral Session 4

Patrice Boucher, Pierrich Plusquellec, Pierre Dufour, Najim Dehak, Patrick Cardinal and Pierre Dumouchel: *PHYSIOSTRESS: A Multimodal Corpus of Data on Acute Stress and Physiological Activation*

Dimitra Anastasiou and Kirsten Bergmann: *A Gesture-Speech Corpus on a Tangible Interface*

Costanza Navarretta: *Filled pauses, Fillers and Familiarity in Spontaneous Conversations*

17:45 – 18:00 – Concluding remarks

Editors

Jens Edlund	KTH Royal Institute of Technology
Dirk Heylen	University of Twente
Patrizia Paggio	University of Copenhagen/University of Malta

Workshop Organizers

Jens Edlund	KTH Royal Institute of Technology
Dirk Heylen	University of Twente
Patrizia Paggio	University of Copenhagen/University of Malta

Table of Contents

Author Index

Stereotyping and Bias in the Flickr30k Dataset

Emiel van Miltenburg
Vrije Universiteit Amsterdam
emiel.van.miltenburg@vu.nl

Abstract

An untested assumption behind the crowdsourced descriptions of the images in the Flickr30k dataset (Young et al., 2014) is that they "focus only on the information that can be obtained from the image alone" (Hodosh et al., 2013, p. 859). This paper presents some evidence against this assumption, and provides a list of biases and unwarranted inferences that can be found in the Flickr30k dataset. Finally, it considers methods to find examples of these, and discusses how we should deal with stereotype-driven descriptions in future applications.

Keywords: image annotation, stereotypes, bias, Flickr30k

1. Introduction

The Flickr30k dataset (Young et al., 2014) is a collection of over 30,000 images with 5 crowdsourced descriptions each. It is commonly used to train and evaluate neural network models that generate image descriptions (e.g. (Vinyals et al., 2015)). An untested assumption behind the dataset is that the descriptions are based on the images, and nothing else. Here are the authors (about the Flickr8k dataset, a subset of Flickr30k):

> "By asking people to describe the people, objects, scenes and activities that are shown in a picture without giving them any further information about the context in which the picture was taken, we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone." (Hodosh et al., 2013, p. 859)

What this assumption overlooks is the amount of *interpretation* or *recontextualization* carried out by the annotators. Let us take a concrete example. Figure 1 shows an image from the Flickr30k dataset.

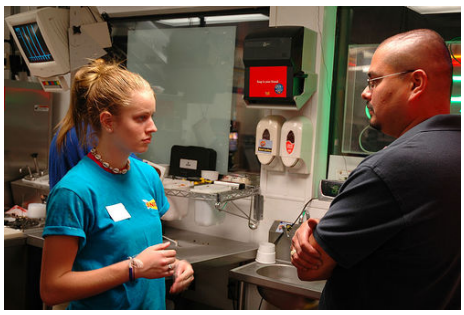

Figure 1: Image 8063007 from the Flickr30k dataset.

This image comes with the five descriptions below. All but the first one contain information that cannot come from the image alone. Relevant parts are highlighted in **bold**:

1. A blond girl and a bald man with his arms crossed are standing inside looking at each other.
2. A **worker** is **being scolded** by her **boss** in a **stern lecture**.
3. A **manager talks to an employee about job performance**.
4. A hot, blond girl **getting criticized by her boss**.
5. Sonic employees **talking about work**.

We need to understand that the descriptions in the Flickr30k dataset are subjective descriptions of events. This can be a good thing: the descriptions tell us what are the salient parts of each image to the average human annotator. So the two humans in figure 1 are relevant, but the two soap dispensers are not. But subjectivity can also result in *stereotypical* descriptions, in this case suggesting that the male is more likely to be the manager, and the female is more likely to be the subordinate. Rashtchian et al. (2010) do note that some descriptions are speculative in nature, which they say hurts the accuracy and the consistency of the descriptions. But the problem is not with the lack of consistency here. Quite the contrary: the problem is that stereotypes are pervasive enough for the data to be consistently biased. And so language models trained on this data may propagate harmful stereotypes, such as the idea that women are less suited for leadership positions.

This paper aims to give an overview of linguistic bias and unwarranted inferences resulting from stereotypes and prejudices. I will build on earlier work on linguistic bias in general (Beukeboom, 2014), providing examples from the Flickr30k data, and present a taxonomy of unwarranted inferences. Finally, I will discuss several methods to analyze the data in order to detect biases.[1]

2. Stereotype-driven descriptions

Stereotypes are ideas about how other (groups of) people commonly behave and what they are likely to do. These ideas guide the way we talk about the world. I distinguish two kinds of verbal behavior that result from stereotypes: (i) linguistic bias, and (ii) unwarranted inferences. The former is discussed in more detail by Beukeboom (2014), who defines linguistic bias as "a systematic asymmetry in word choice as a function of the social category to which the target belongs." So this bias becomes visible through the *distribution* of terms used to describe entities in a particular

[1] The Flickr30k data also contains examples where annotators judge the subjects of the images on their looks. E.g. description #4 above calling the girl in the image *hot*. Analyzing this judgmental language goes beyond the scope of this paper.

category. Unwarranted inferences are the result of speculation about the image; here, the annotator goes beyond what can be glanced from the image and makes use of their knowledge and expectations about the world to provide an overly specific description. Such descriptions are directly identifiable as such, and in fact we have already seen four of them (descriptions 2–5) discussed earlier.

2.1. Linguistic bias

Generally speaking, people tend to use more concrete or specific language when they have to describe a person that does not meet their expectations. Beukeboom (2014) lists several linguistic 'tools' that people use to mark individuals who deviate from the norm. I will mention two of them.[2]

Adjectives One well-studied example (Stahlberg et al., 2007; Romaine, 2001) is sexist language, where the sex of a person tends to be mentioned more frequently if their role or occupation is inconsistent with 'traditional' gender roles (e.g. *female surgeon, male nurse*). Beukeboom also notes that adjectives are used to create "more narrow labels [or subtypes] for individuals who do not fit with general social category expectations" (p. 3). E.g. *tough woman* makes an exception to the 'rule' that women aren't considered to be tough.

Negation can be used when prior beliefs about a particular social category are violated, e.g. *The garbage man was not stupid.* See also (Beukeboom et al., 2010).

These examples are similar in that the speaker has to put in additional effort to mark the subject for being unusual. But they differ in what *we* can conclude about the speaker, especially in the context of the Flickr30k data. Negations are much more overtly displaying the annotator's prior beliefs. When one annotator writes that *A little boy is eating pie **without** utensils* (image 2659046789), this immediately reveals the annotator's normative beliefs about the world: pie should be eaten *with* utensils. But when another annotator talks about *a girls basketball game* (image 8245366095), this cannot be taken as an indication that the annotator is biased about the gender of basketball players; they might just be helpful by providing a detailed description. In section 3 I will discuss how to establish whether or not there is any bias in the data regarding the use of adjectives.

2.2. Unwarranted inferences

Unwarranted inferences are statements about the subject(s) of an image that go beyond what the visual data alone can tell us. They are based on additional assumptions about the world. After inspecting a subset of the Flickr30k data, I have grouped these inferences into six categories (image examples between parentheses):

Activity We've seen an example of this in the introduction, where the 'manager' was said to be *talking about job performance* and *scolding [a worker] in a stern lecture* (8063007).

[2]Examples given are also due to (Beukeboom, 2014).

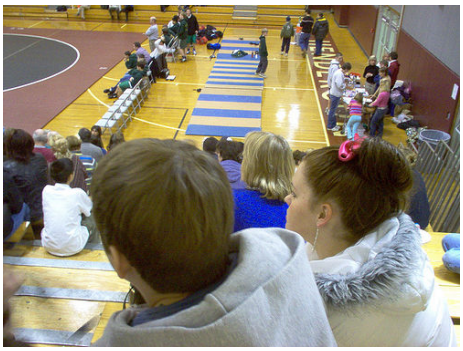

Figure 2: Image 4183120 from the Flickr30k dataset.

Ethnicity Many dark-skinned individuals are called *African-American* regardless of whether the picture has been taken in the USA or not (4280272). And people who look Asian are called Chinese (1434151732) or Japanese (4834664666).

Event In image 4183120 (figure 2), people sitting at a gym are said to be watching a game, even though there could be any sort of event going on. But since the location is so strongly associated with sports, crowdworkers readily make the assumption.

Goal Quite a few annotations focus on explaining the *why* of the situation. For example, in #3963038375 a man is fastening his climbing harness *in order to have some fun*. And in an extreme case, one annotator writes about a picture of a dancing woman that *the school is having a special event in order to show the american culture on how other cultures are dealt with in parties* (3636329461). This is reminiscent of the Stereotypic Explanatory Bias (Sekaquaptewa et al., 2003, SEB), which refers to "the tendency to provide relatively more explanations in descriptions of stereotype inconsistent, compared to consistent behavior" (Beukeboom et al., 2010, p. 5). So in theory, odd or surprising situations should receive more explanations, since a description alone may not make enough sense in those cases, but it is beyond the scope of this paper to test whether or not the Flickr30k data suffers from the SEB.

Relation Older people with children around them are commonly seen as parents (5287405), small children as siblings (205842), men and women as lovers (4429660), groups of young people as friends (36979).

Status/occupation Annotators will often guess the status or occupation of people in an image. Sometimes these guesses are relatively general (e.g. college-aged people being called *students* in #36979), but other times these are very specific (e.g. a man in a workshop being called a *graphics designer*, 5867606).

3. Detecting stereotype-driven descriptions

In order to get an idea of the kinds of stereotype-driven descriptions that are in the Flickr30k dataset, I made a browser-based annotation tool that shows both the images

Asian		Average 60%
2339632913	Asian child/baby	2
3208987435	Asian baby, Asian/oriental woman	3
7327356514	Asian girl/baby, Asian/oriental woman	4
Black		Average 40%
1319788022	African-American (AA)/black baby	3
149057633	African/AA child, black baby	3
3217909454	Dark-skinned baby	1
3614582606	AA baby	1
White		Average 20%
11034843	White baby boy	1
176230509	White baby boy	1
2058947638	White baby	1
3991342877	White baby	1
4592281294	White baby stroller	FP
661546153	White baby stroller	FP
442983801	Fair-skinned baby	1

Table 1: Number of times ethnicity/race was mentioned per category, per image. The average is expressed as a percentage of the number of descriptions. Counts in the last column correspond to the number of descriptions containing an ethnic/racial marker. Images were found by looking for descriptions matching `(asian|white|black|African-American|skinned) baby`. I found two false positives, indicated with FP.

and their associated descriptions.[3] You can simply leaf through the images by clicking 'Next' or 'Random' until you find an interesting pattern.

3.1. Ethnicity/race

One interesting pattern is that the ethnicity/race of babies doesn't seem to be mentioned *unless* the baby is black or asian. In other words: white seems to be the default, and others seem to be marked. How can we tell whether or not the data is actually biased?

We don't know whether or not an entity belongs to a particular social class (in this case: ethnic group) until it is marked as such. But we can approximate the proportion by looking at all the images where the annotators have used a marker (in this case: adjectives like *black, white, asian*), and for those images count how many descriptions (out of five) contain a marker. This gives us an *upper bound* that tells us how often ethnicity is indicated by the annotators. Note that this upper bound lies somewhere between 20% (one description) and 100% (5 descriptions). Table 1 presents count data for the ethnic marking of babies. It includes two false positives (talking about a *white baby stroller* rather than a *white baby*). In the Asian group there is an additional complication: sometimes the mother gets marked rather than the baby. E.g. *An Asian woman holds a baby girl.* I have counted these occurrences as well.

The numbers in table 1 are striking: there seems to be a real, systematic difference in ethnicity marking between the groups. We can take one step further and look at all the 697 pictures with the word 'baby' in it. If there turn out to be disproportionately many white babies, this strengthens the conclusion that the dataset is biased.[4]

I have categorized each of the baby images. There are 504 white, 66 asian, and 36 black babies. 73 images do not contain a baby, and 18 images do not fall into any of the other categories. While this does bring down the average number of times each category was marked, it also increases the contrast between white babies (who get marked in less than 1% of the images) and asian/black babies (who get marked much more often). A next step would be to see whether these observations also hold for other age groups, i.e. children and adults.[3]

3.2. Other methods

It may be difficult to spot patterns by just looking at a collection of images. Another method is to tag all descriptions with part-of-speech information, so that it becomes possible to see e.g. which adjectives are most commonly used for particular nouns. One method readers may find particularly useful is to leverage the structure of Flickr30kEntities (Plummer et al., 2015). This dataset enriches Flickr30k by adding coreference annotations, i.e. which phrase in each description refers to the same entity in the corresponding image. I have used this data to create a coreference graph by linking all phrases that refer to the same entity. Following this, I applied Louvain clustering (Blondel et al., 2008) to the coreference graph, resulting in clusters of expressions that refer to similar entities. Looking at those clusters helps to get a sense of the enormous variation in referring expressions. To get an idea of the richness of this data, here is a small sample of the phrases used to describe beards (cluster 268): *a scruffy beard; a thick beard; large white beard; a bubble beard; red facial hair; a braided beard; a flaming red beard.* In this case, 'red facial hair' really stands out as a description; why not choose the simpler 'beard' instead?[5]

4. Discussion

In the previous section, I have outlined several methods to manually detect stereotypes, biases, and odd phrases. Because there are infinitely many ways in which a phrase can be biased, it is almost impossible to remove this bias from the data. So how should we deal with stereotype-driven descriptions?

Neutralizing stereotypes for production One way to move forward might be to work with multilingual data. Elliott et al. (2015) propose a model that generates image descriptions given data from multiple languages, in their case German and English. Multilingual, or better: multicultural data might force models to put less emphasis on features that are only salient to annotators from one particular country.

Stereotypes and interpretation While stereotypes might be a problem for production, further study of cultural

[3] Code and data is available on GitHub: `https://github.com/evanmiltenburg/Flickr30k-Image-Viewer`

[4] Of course this extra step does constitute an additional annotation effort, and it is fairly difficult to automate; one would have to train a classifier for each group that needs to be checked.

[5] Code and data is available on GitHub: `https://github.com/evanmiltenburg/Flickr30k-clusters`

stereotyping might be beneficial to systems that have to interpret human descriptions and determine likely referents of those descriptions. E.g. knowing that *baseball player* probably refers to a *male* baseball player is very useful.

Levels of describing an image There is a large body of work in art, information science, library science and related fields dedicated to the description and categorization of images (Shatford, 1986; Jaimes and Chang, 1999). A common thread is that we can divide image description into multiple levels or stages, starting from concrete physical attributes up to abstract contextual information. These levels build on each other; we first have to recognize separate entities before we can reason about their relation. But recent neural network models like (Vinyals et al., 2015) do not match this procedure. Rather, they are trained to create a direct mapping between images and their descriptions. With this paper, I hope to have shown that the Flickr30k dataset is *layered*, reflecting not only the physical contents of the images, but also whether the images match the everyday expectations of the annotators. An interesting challenge would be for image description models to learn separate representations for both layers: the perceptual and the contextual.

5. Conclusion

This paper provided a taxonomy of stereotype-driven descriptions in the Flickr30k dataset. I have divided these descriptions into two classes: linguistic bias and unwarranted inferences. The former corresponds to the annotators' choice of words when confronted with an image that may or may not match their stereotypical expectancies. The latter corresponds to the tendency of annotators to go beyond what the physical data can tell us, and expand their descriptions based on their past experiences and knowledge of the world. Acknowledging these phenomena is important, because on the one hand it helps us think about what is learnable from the data, and on the other hand it serves as a warning: if we train and evaluate language models on this data, we are effectively teaching them to be biased.

I have also looked at methods to detect stereotype-driven descriptions, but due to the richness of language it is difficult to find an automated measure. Depending on whether your goal is production or interpretation, it may either be useful to suppress or to emphasize biases in human language. Finally, I have discussed stereotyping behavior as the addition of a contextual layer on top of a more basic description. This raises the question what kind of descriptions we would like our models to produce.

6. Acknowledgments

Thanks to Piek Vossen and Antske Fokkens for discussion, and to Desmond Elliott for comments on an earlier version of this paper. This research was supported by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019).

7. Bibliographical references

Beukeboom, C. J., Finkenauer, C., and Wigboldus, D. H. (2010). The negation bias: when negations signal stereotypic expectancies. *Journal of personality and social psychology*, 99(6):978.

Beukeboom, C. J. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In J. Laszlo, et al., editors, *Social cognition and communication*, volume 31, pages 313–330. Psychology Press. Author's pdf: `http://dare.ubvu.vu.nl/handle/1871/47698`.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Elliott, D., Frank, S., and Hasler, E. (2015). Multilingual image description with neural sequence models. *CoRR*, abs/1510.04709.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899.

Jaimes, A. and Chang, S.-F. (1999). Conceptual framework for indexing visual information at multiple levels. In *Electronic Imaging*, pages 2–15. International Society for Optics and Photonics.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Romaine, S. (2001). A corpus-based view of gender in british and american english. *Gender Across Languages: The linguistic representation of women and men*, 1:153–175.

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., and von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*, 39(1):75–82.

Shatford, S. (1986). Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly*, 6(3):39–62.

Stahlberg, D., Braun, F., Irmen, L., and Sczesny, S. (2007). Representation of the sexes in language. *Social communication*, pages 163–187.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

The Multimodal HuComTech Corpus: Principles of Annotation and Discovery of Hidden Patterns of Behaviour

Laszlo Hunyadi, Tamas Varadi, Istvan Szekrenyes
University of Debrecen, Hungarian Academy of Science, University of Debrecen
H-4032 Debrecen, Egyetem tér 1., H-1068 Budapest VI., Benczúr u. 33., H-4032 Debrecen, Egyetem tér 1.
hunyadi@unideb.hu, varadi.tamas@nytud.mta.hu, szekrenyes.istvan@arts.unideb.hu

Abstract

The understanding of behaviour, and human behaviour in particular, involves several challenges. One of the probably most important challenges is that, due to the multimodal nature of interactions, one and the same communicative function can be expressed variably by a single verbal or nonverbal event or their combination, and in such a way that in many cases none of these functional markers is mandatory. The obvious question then arises, how, by following what mechanism and observing what markers and their structural relations is the understanding of a given behaviour made possible. The multimodal HuComTech corpus was built by assuming as behavioural markers a large set of (a) formally characterizable physical events, such as gaze, posture, head and hand movement as well as prosody, and (b) markers resulting from the interpretation of certain configurations, such as perceived emotions, discourse and pragmatic events. In order to actually understand a given behaviour through specifying the general characteristics of patterns associated with behavioural functions, the existing large amount of data within the corpus are processed applying descriptive statistics as well as more advanced heuristics, the latter by both machine learning and using the framework Theme dedicated for the discovery of hidden patterns of behaviours.

Keywords: HuComTech, multimodality, annotation, hidden patterns

1. Introduction: About the corpus

The HuComTech Corpus represents a detailed and extensive annotation of verbal and nonverbal human behaviour as manifested in formal and informal dialogues of more than 50 hours of audiovisual recordings. The participants were 110 university students, and the language of the dialogues in both settings was Hungarian. The initial aim of building the corpus was to acquire a wide range of data characteristic of human-human interaction in order to make generalisations for their implementation in more advanced human-machine interaction systems (Hunyadi, 2011). We were especially interested in the formal and pragmatic ways of how dialogues are managed according to specific contexts (Hunyadi et al., 2016). The view is becoming increasingly shared that, in order to make a conversation successful, the formal verbal, syntactic and semantic aspects of a conversation need to go hand in hand with its nonverbal aspects (the suprasegmentals of speech as well as a wide variety of gestures). It is especially important in a human-machine interaction, where a proper emphasis on multimodality can significantly add to the robustness of such systems: the recognition of a speaker's gestures by the machine agent can contribute to the generation of its contextually proper responses and, consequently, to the sense of cooperativeness, a crucial component of a successful interaction (Vilhjalmsson et al., 2007). The need to cooperate goes beyond understanding the propositional content of the verbal component that is enhanced by the gestural one: the participants (either human or machine) need to interpret the partner's intentions, as well as his/her emotions as complements to their actions and manifestations of reactions to such intentions (McNeill, 1992; Enfield, 2009). Therefore the annotation of the corpus was extended to those formal and pragmatic markers (Allwood et al., 2007) of the given dialogues that were considered both characteristic of human-human interactions and implementable in a human-machine interaction system.

2. The annotation of the corpus: its layers and attributes

With these fundamental aims in mind the corpus was annotated both for video and audio, and at both levels both for their formal and pragmatic properties (Ágnes Abuczki and Esfandiari-Baiat, 2013). As for its formal aspects, those properties were annotated which could be described in terms of physical attributes. Whereas they were done predominantly manually, some (especially prosody) was annotated or evaluated using automatic algorithms.

As a special feature of the corpus, annotation was done, when applicable, both multimodally and unimodally. The rationale behind it was that whereas it is generally accepted that both the production and the perception/interpretation of a communicative event is essentially multimodal due to the participation of a number of (verbal and non-verbal) channels (modalities), both the analysis and generation of such an event by the machine agent needs to follow a complex of individual modalities, i.e. by the setting of the parameters of each of the modalities separately.

The annotation scheme of the individual levels is summarized in the following tables (Table 1 – 6). The "audio annotation" (see Table 1) was the first scheme in the corpus including various attributes (textual transcriptions, emotions, discourse structure etc.) which are annotated based on the audio signal using intonation phrases (head and subordinate clauses) as segmentation unit. The "syntactic annotation" (see Table 2) aims to segment the textual transcriptions into sentences and clauses analysing their syntactic relations and missing elements (Hunyadi et al., 2012). The "video annotation" (see Table 2) contains the annotation of non-verbal gestures and perceivable emotions (Ekman and

Friesen, 1969) using the video signal and a custom annotation tool, the Qannot program (Pápay et al., 2011) which was developed at the University of Debrecen. For the automatic annotation of facial expressions, we also had an experiment with the *FaceReader* application from *Noldus*.[1]

Audio	
theoretical model	descriptive
modalities	verbal non-verbal acoustic
annotation mode	manual
validation mode	semi-automatic
annotation tool	Praat
related ISO categories	partly available
annotated elements	transcription fluency intonation phrases iteration embeddings emotions turn management discourse

Table 1: Audio annotation

Syntax	
theoretical model	descriptive
modalities	verbal
annotation mode	manual
validation mode	semi-automatic
annotation tool	Praat
related ISO categories	not available
annotated elements	sentences clauses POS (in progress) hierarchical organization missing elements

Table 2: Syntactic annotation

The "pragmatic annotation" has two schemes: the unimodal (see Table 4) and the multimodal (see Table 5) pragmatic annotation. In case of the unimodal annotation, the annotators cannot hear the audio, therefore the verbal content of the interactions was not available.

[1] `http://www.noldus.com/human-behavior-research/products/facereader`

Video	
theoretical model	Ekman & Friesan: emotions, emblems
	descriptive: other tiers
modalities	visual
annotation mode	manual automatic
validation mode	manual
annotation tool	Qannot
related ISO categories	partly available
annotated elements	facial expressions gaze eyebrows headshift handshape touchmotion posture deixis emblem emotions

Table 3: Video annotation

Unimodal pragmatics	
theoretical model	modified (single-modal) version of conversation analysis
modalities	visual
annotation mode	manual
validation mode	manual
annotation tool	Qannot
related ISO categories	partly available
annotated elements	turn management attention agreement deixis information structure

Table 4: Unimodal pragmatic annotation

Multimodal pragmatics	
theoretical model	modified (multimodal) version of Speech Act Theory
modalities	visual verbal
annotation mode	manual
validation mode	manual
annotation tool	Qannot
related ISO categories	partly available
annotated elements	communicative acts supporting acts thematic control information structure

Table 5: Multimodal pragmatic annotation

Prosody	
theoretical model	psycho-acoustic model of tonal perception
modalities	non-verbal acoustic
annotation mode	automatic
validation mode	N/A
annotation tool	Praat
related ISO categories	not available
annotated elements	pitch intensity pauses speech rate

Table 6: Prosodic annotation

3. Automatic annotation of prosody

Most of the annotation labels were created manually based on the observation of well-trained annotators. The only exception was prosody, which was annotated by a computer algorithm (Szekrenyes, 2014) using the built-in scripting language of Praat Speech Processing Tool (Boersma and Weenik, 2016). The development aims at making an annotation procedure which can be generally used for the prosodic analysis of any spoken language corpora containing the audio of two-party interactions. Therefore not only the annotation results but also its methodology could be important to highlight as a new language resource used in the HuComTech-corpus. Following the work of Piet Mertens (Mertens, 2004), the purpose of the annotation is to provide a psycho-acoustically relevant representation of prosodic features including intonation, intensity and speech rate. The main difference is that the prosodic segmentation does not follow the syllable-size units of Merten's Prosogram, but one event can integrate sequence of syllables in larger trends of modulation which are classified based on dynamic, speaker-dependent thresholds (instead of glissando). Contrary to other existing tools, it is language independent (unlike ToBi) and does not require any training material: only a two-level annotation of the speaker change. The output of the final algorithm (Szekrényes 2016) contains annotated segments of prosodic events describing larger, smoothed and stylized movements of the originally measured data (F0 and intensity values), where the labels indicate the shape (descending, falling, rising etc.), the absolute (in hertz, decibel or syllable/second) and the relative (adjusting to the individual characteristics of the speaker) vertical position of every single prosodic event through their starting and ending points. The resulting labels represent these two-dimensional (modulations and positions) prosodic structure of interactions, which can be considered as an automatically generated, but perceptually verifiable music sheet of communication based on the raw F0 and intensity data.

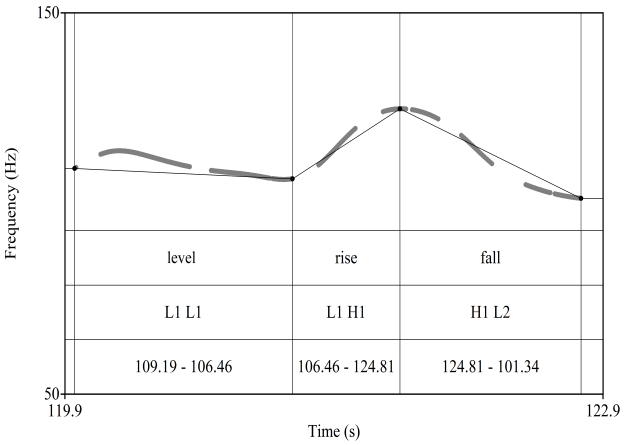

Figure 1: Annotation of intonation

4. The understanding of the corpus: towards the discovery of hidden patterns of behaviour

There are at least two challenges regarding corpus building: what data to collect, and what to learn (what conclusion to draw) from the data collected afterwards. In the best practice, both challenges are based on and reconciled by both a practical goal and a theoretical approach supporting this goal. In the case of the HuComTech Corpus a practical goal was defined at the outset (enhancing HCI by obtaining a more detailed knowledge about relevant aspects of HHI). The greater challenge is, however, the task to make generalisations from more than a million annotation items suitable for this particular knowledge transfer. At present, we are experimenting with two approaches for such generalisations: machine learning testing, among others, the Hidden

Markov Model, and applying a statistical model particularly designed for the discovery of hidden patterns of behaviour in social interactions using the framework Theme (Magnusson, 2000). The talk will put special emphasis on the latter at the core of which is the assumption that patterns of social interactions can be captured as both contiguous and noncontiguous sequences of all sorts of obligatory or optional events happening across an essentially unlimited stretch of time. As an example, the statistical analysis of the data shows an interesting sequence of interactions: within a 10 minute formal dialogue between a female agent A and a male speaker B there were 36 occurrences of a pattern with B showing the visual signs of an intention to speak followed by B breaking in, the latter again followed by B intending to speak. This sequence, supported by its strong statistical significance ($p < 0.000005$) demonstrates the characteristic dynamics of the given interaction between A and B. Following the same sequence across the large number of different recordings in the corpus one can observe the individual differences between the various participants and, ultimately, one can subcategorize the various individual behaviours relevant to and useful for the recognition and implementation of speaker accommodation in the various human-human and human-machine settings.

5. The availability of the HuComTech Corpus

Our intention is to make the corpus available in the coming months as a multimodal language resource for the LREC community. Until then, certain basic annotations are already available at The Language Archive under Donated Corpora[2] as well as at the META-SHARE website. The video recordings are open for research purposes.

6. Bibliographical References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Proceedings of the International Language Resources and Evaluation Conference (LREC)*, 41(3-4):273–287.

Boersma, P. and Weenik, D. (2016). Praat: doing phonetics by computer version 6.0.13. http://www.praat.org.

Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica*, 1(1):49–98.

Enfield, N. (2009). *The Anatomy of Meaning. Speech, gesture, and composite utterances*. Cambridge University Press.

Ágnes Abuczki and Esfandiari-Baiat, G. (2013). An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, 9:86–98.

Hunyadi, L., Szekrényes, I., Borbély, A., and Kiss, H. (2012). Annotation of spoken syntax in relation to prosody and multimodal pragmatics. In *Proceedings of IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 537–541, Kosice, Slovakia, december.

Hunyadi, L., Kiss, H., and Szekrenyes, I., (2016). *Incompleteness and Fragmentation: Possible Formal Cues to Cognitive Processes Behind Spoken Utterances.*, pages 231–257. Springer International Publishing.

Hunyadi, L. (2011). Multimodal human-computer interaction technologies. theoretical modeling and application in speech processing. *Argumentum*, 7:240–260.

Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection behavior research methods. *Behavior Research Methods, Instruments, & Computers*, 32:93–110.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press.

Mertens, P. (2004). Prosogram: semiautomatic transcription of prosody based on a tonal perception model. In *Proceedings of the 2nd International Conference of Speech Prosody*, pages 549–552, Nara, Japan, march.

Pápay, K., Szeghalmy, S., and Szekrényes, I. (2011). Hucomtech multimodal corpus annotation. *Argumentum*, 7:330–347.

Szekrenyes, I. (2014). Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. *Journal on Multimodal User Interfaces*, 8:(2):143–150.

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N., Kipp, M., and Kopp, S. (2007). The behavior markup language: Recent developments and challenges. In *Proceedings of Intelligent Virtual Agents (IVA 2007)*, pages 99–111. Springer, september.

[2] https://hdl.handle.net/1839/00-0000-0000-001A-E17C-1@view

The Annotation of Gesture Designed for Classroom Interaction

Michael Amory and Olesya Kisselev
The Pennsylvania State University
304 Sparks Building, University Park, PA
mda5004@psu.edu; ovk103@psu.edu

Abstract

In the past decade, the field of Applied Linguistics has witnessed an increased interest in the study of multimodal aspects of language and language acquisition, and the number of multimodal corpora that are designed to investigate classroom interactions, second language acquisition and second language pedagogy is on the rise. The promise is that these digital repositories of video-recordings will be able to take advantage of Corpus Linguistics tools and procedures in order to maximize and diversify analytical capabilities. However, the transcription conventions (i.e., annotation schemas) for multimodal features (such as gestures, gaze, and body movement) that are simple, systematic and searchable are not readily available. The current project focuses on developing an annotation schema for the transcription of gestures, integrating the research traditions Conversation Analysis, gesture research, ASL and Corpus Linguistics. The goal of the project is to create a set of conventions that have analytical and descriptive power required for gesture research but are manageable for the transcriber and reader to engage with and that are systematic to allow for searchability. The study utilizes video-recorded data from the Corpus of English for Academic and Professional Purposes developed at the Pennsylvania State University.

Keywords: Conversation Analysis, Multimodal Corpus Linguistics, Gesture Annotation

In the last decade, the field of Applied Linguistics has been witness to a rise in development of corpora designed to investigate classroom interactions and second language (L2) pedagogy (e.g., Lab School at Portland State University, Reder, 2005). These projects have enabled researchers and language teaching practitioners to conduct cross-case, cross-corpora comparisons on various interactional practices in the classroom environment. One such video-based specialized corpus is the Corpus of English for Academic and Professional Purposes (CEAPP) at the Pennsylvania State University (the focus and testing ground of the present project). In the hopes of creating an important knowledge base upon which practitioners and researchers can draw to identify problems, devise solutions, and enhance efficacy in classroom interactions, CEAPP video-recordings are transcribed using Conversation Analysis (CA) conventions (cf. Jefferson, 2004) and are subjected to CA analyses.

CEAPP functions as a digital repository consisting of approximately 350 hours of classroom interactions that provides corpus resources focusing on the teaching and learning of ESL, as well as New Professoriate Initiatives (NPI) focusing on Science, Technology, Engineering, and Mathematics (STEM) courses. CEAPP does provide basic search capabilities as one can use the search interface to search for transcripts of classroom interactions from a variety of courses and language proficiency levels. For example, one can specify a single or a combination of search criteria, including course type, course level, teaching context, activity type, professor rank, professor education, professor experience, etc. Although CEAPP may be considered a corpus (i.e., a principled collection of data), it does not possess most of the functionalities of a searchable corpus. We believe, however, that CEAPP will be a significantly more powerful research platform if it combined with CL capabilities, both at micro-level (i.e., linguistic structures and utterances) and macro-level (i.e., discourse) (Walsh, 2013). The ability to annotate data with tags based on CA conventions and other multimodal features and then search the corpus by these tags will significantly enhance a researcher's engagement and profound understanding of data. Of particular interest to the current stage of the project, is the annotation of multimodal components, specifically gestures, which has proven to be a challenging endeavor for both CA and CL.

The study of multimodality has become an area of increasing research interest in the recent decades; recent studies of gesture have created a considerable body of supporting evidence for language's close relationship to bodily movement and argue that gesture and speech are part of a unified system and should not be analyzed separately (McNeill, 1992; Goldin-Meadow, 2005). In these studies, several gesture classification systems have been proposed (i.e., Ekman & Friesen, 1969; Freedman & Hoffman, 1967; McNeill, 1992). However, since these systems were designed with particular research questions in mind, they may not be immediately applicable to large databases created and designed for CA or CL research. Consider, for example, a sample transcript in Illustration 1 below (from Stam, 2014).

(3) [[/ and] [/ / go down the pipe all the way] [/ to the street]]
a b c

- a: iconic: both hands, facing center, fingers facing away on both sides of the body on right and left extreme periphery <Sylvester with the ball inside of his stomach>;
- b: iconic: both hands, facing center, fingers away from body, right hand at right center periphery, left had at upper left periphery move down to the right across body to low right periphery and flip up <Sylvester + balling bowl going down and out the pipe> PATH;
- c: deictic: right hand turns over and points down at low right periphery palm towards center fingers toward down, left hand lowers to right center palm towards body, fingers toward right and both hands hold <location of street + endpoint>.

Illustration 1: Gesture annotation from Stam (2014)

It is not our intent to critique existing methodologies as they have their own analytic purpose and have contributed significant insights to their respective field. However, they are challenging for the purposes of CA and CL in a number of important ways. Most importantly, annotations as the one presented above usually include a (lesser or higher) degree of interpretation made by the researcher (notice the use of such semantic categories as "iconic" and "deictic" in Stam's transcript); these interpretations vary across studies and research traditions, and may change across time, leading to the lack of systematicity that then leads to issues with searchability. That is, with the lack of a unified, descriptive system to describe the gestures themselves, their location and movement in relation to the body, etc., it becomes more challenging to search across multiple transcripts. Another issue impacting searchability is the lack of simplicity within the transcript: gesture annotation can be difficult for the non-expert reader to understand and for a transcriber to record systematically (on top of requiring an increasing number of transcription hours).

CEAPP attempts to create a system that is based on the previous gesture research but is grounded in the tradition of Jefferson (1974), a CA analyst who wished to create a methodology to annotate transcripts of spoken interaction that was a compromise between two objectives: to preserve the details of talk (in this case, gesture) as it is actually produced (description before interpretation), while at the same time remaining simple enough to yield transcripts that are accessible to a general audience (simplicity). In addition, the current project will attempt to take this one step further and offer a systematic transcription of gesture designed for both CA and CL research while taking into account the large database such CEAPP (searchability). Thus, a balanced must be reached between the simplicity and readability of multimodal tags through the lens of CA and their searchability and analytic power in a large corpus.

To satisfy the requirements of descriptive power, systematicity and simplicity, we draw upon previous research in gesture, particularly McNeill (1992), as well as adopt parameters and classifications from American Sign Language research and sign language phonology (Valli & Lucas, 2000). The parameters that are considered to be pertinent are the following: *handshape, movement, palm orientation, and movement*. In addition, we have added handedness for descriptive purposes. Some parameters are more elaborated upon (e.g., handshape, movement) since previous research has linked handshape and type of movement to cognitive-linguistic categories (e.g. type of movement corresponds with linguistic categories of motion such as PATH and MANNER, see for example Stam, 2008; Cadierno, 2010).

What follows is by no means an exhaustive list of the gesture annotation system. Rather, we present a broad overview of each category with select, representative examples and classifications.

1. Handedness indicates which hand is used in gesturing.

Right Hand (RH)
Left Hand (LH)
Both Hands (BH)

2. Handshape description is based on the complex visual-spatial system (Nakamura, 2002) used in American Sign Language (ASL). Handshape is literally the shape (or shapes) in which we form our hand during the production of a gesture (i.e., hand configuration). The utilization of this system presents us with a purely descriptive account to represent handshapes while trying to avoid implying meaning. In illustrations 2-5 below, a few examples of handshapes are presented. For reference, handshape is indicated by HS. The number or letter follow HS represents the form that the hand has taken. For reference, these numbers or letters are based upon ASL.

Illustration 2: HS-1

Illustration 3: HS-V

Illustration 4: HS-O

3. Orientation of the palm while making the handshape.

Left (L)
Right (R)
Up (U)
Down (D)
Front (F)
Back (B)

4. Location of handshape in relation to the body. This parameter is a modified and significantly simplified account of McNeill's 1992 original proposition, which suggested 21 different locales around in space around the body (such as Extreme Periphery, Lower Left, Center-Center, etc.)

Center (C)
Left (L)
Right (R)
Upper left (UL)
Upper right (UR)
Lower left (LL)
Lower right (LR)

5. Movement of the handshape in relation to body. This parameter will be used to represent the trajectory and type of movement.

Single movement (SM)
Repetitive movement (RM), 3x, 2x, etc.
Clockwise (CW)
Counter-clockwise (CCW)
Sinuous (SI)
Straight (ST)
Wrist rotation/movement (WR)

Figure 1 below is an illustration of the gesture annotated alongside the co-occurring utterance in Excerpt 1. The gesture made in Figure 1 is a HS-bent-5. As an example of how to tag a gesture using CL and the annotation schema is presented below Figure 1.

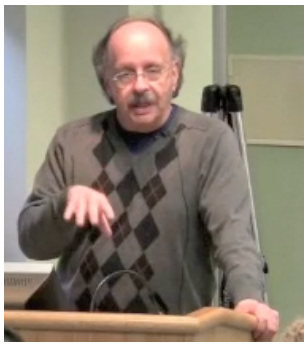

Figure 1: HS-bent-5

Tag: _RH_HS-bent-5_PD_C_SM

Excerpt 1 below represents how the annotation of gesture may be incorporated into a CA transcript using the annotation schema.

Excerpt 1: Example of CA transcript

8	*TEA:	{>th- the< gestures show up,
9		{RH HS-bent-5 moves from R to C
10	*TEA:	{in interesting wa:yz.
11		{RH HS-bent-5 moves from C to R

Conclusion

With the enhancement of technology, digital repositories such as CEAPP will be able to take advantage of CL tools and procedures in order to maximize and diversify analytical capabilities. Utilizing coding schemes for the annotation of multimodal corpora, in our case tagging of gestures and incorporation of searchable functions, may facilitate cross-case studies, cross-corpora, and longitudinal analyses. At the same time, CA methodology has a lot to offer to corpora studies of speech and communication, especially in terms of accounting for multimodality of naturally-occurring speech. With a commitment to "naturalistic inquiry" (Schegloff, 1997, p. 501) and rigorous transcription procedures, CA can provide a theoretical and practical framework to transcribing and analyzing video recordings along with transcripts of these recordings in a systematic, simple and yet descriptive way. This would allow researchers to maintain a clear distinction between form and function in gesture transcription and annotation. By combining CA and CL approaches to data transcription, coding and analysis, can reveal new insights into the relationship between interaction patterns, language use, and learning (O'Keeffe & Walsh, 2012; Walsh, 2012).

Bibliographical References

Bucholtz, M. (2000). The politics of transcription. *Journal of pragmatics*, *32*(10), pp. 1439-1465.

Cadierno, T. (2010). Motion in Danish as a second language: Does the learner's L1 make a difference? In Z. Han & T. Cadierno (Eds.), *Linguistic Relativity in SLA* (pp. 1-33). Bristol: Multilingual Matters.

Ekman, P. & Friesen, W.V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotic, 1*, pp. 49-98.

Freedman, N., & Hoffman, S. P. 1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor Skills*, 24, pp. 527-539.

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

Jefferson, G. (1974). Error correction as an interactional resource. *Language in Society*, *3*(2), pp. 181-199.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.) *Conversation Analysis: Studies from the first generation* (pp. 13-23). Philadelphia: John Benjamins.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Nakamura, K. (2002). Sign Language Linguistics. About American Sign Language. Deaf Resource Library. Retrieved 12, Feb. 2016 <http://www.deaflibrary.org/asl.html>.

O'Keeffe, A., & Walsh, S. (2012). Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory*, *8*(1), pp. 159-181.

Reder, S. (2005). The "lab school.". *Focus on Basics*, *8*, pp. 1-7.

Schegloff, E. A. (1997). Whose text? Whose context?. *Discourse & Society*, *8*(2), pp. 165-187.

Stam, G. (2008). What gestures reveal about second language acquisition. *Gesture: Second Language Acquisition and Classroom Research*, 231.

Stam, G. (2014). Further changes in L2 thinking for speaking? In C. Mller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and S. Tessendorf (Eds.) *Body Language Communication: An international handbook on multimodality in human interaction*. *Volume 2*, 1875-1886. Handbooks of Linguistics and CommunicationScience/Handbcher zur Sprach- und Kommunikationswissenschaft. Berlin, New York: Mouton De Gruyter.

Valli, C., & Lucas, C. (2000). *Linguistics of American sign language: An introduction*. Gallaudet University Press.

Walsh, S. (2013). Corpus linguistics and conversation analysis at the interface: Theoretical perspectives, practical outcomes. In *Yearbook of Corpus Linguistics and Pragmatics 2013* (pp. 37-51). Springer Netherlands.

“Do You Like a Cup of Coffee?” - The CASIA Coffee House Corpus

Minghao Yang*, Ronald Böck†,*, Dawei Zhang*, Tingli Gao*, Linlin Chao*, Hao Li*, Jianhua Tao*

* National Laboratory of Pattern Recognition, Chinese Academy of Sciences Beijing,
Zhongguancun East Rd. 96, Beijing 100190, China

† Cognitive Systems Group, Otto von Guericke University Magdeburg,
Universitätsplatz 2, 39106 Magdeburg, Germany

mhyang@nlpr.ia.ac.cn, ronald.boeck@ovgu.de

Abstract

Virtual agents are perfect options to represent a technical device in an interaction. However an appearance may influence in particular the user’s behaviour. Therefore, we present a corpus containing naturalistic communication between a user and two virtual agents in a coffee house environment. The scenario is set up in a pleasant way dealing with various topics like discussion on coffee, drinks, weather, and gaming. Thus, behavioural studies are feasible. The combination of the agents is inspired by the “Utah scenario” but fitting two screens which allows for a non-static behaviour of the user. The multimodally recorded dataset provides audio-visual material in Mandarin from more than 50 participants. Further, Kinect recordings and transcripts are available as well. Besides the corpus description first insights on the material are given. Based on the Kinect’s data movement and behaviour analyses of users in a Human-Computer Interaction are possible. Currently we identified two prototypical movement patterns in this context. Furthermore, we investigated in parallel the user’s movement, head direction information, and the dialogue course. An interesting finding is that in such an interaction a full turn of the head and body happens relatively late even if the turn already fully shifted between agents.

Keywords: Corpus Description, Virtual Agents, Human-Computer Interaction, Behaviour and Dialogue Analysis

1. Introduction

Natural human-like talking avatars, as an exciting modality for Human-Computer Interaction (HCI), have been studied intensively for the last ten years. Since the pioneering work on artificial agents remarkable progress has been achieved in human-like avatars (cf. (Cassell et al., 1994; Cerekovic et al., 2009; Courgeon et al., 2010; Kipp et al., 2010; van Welbergen et al., 2010). With the improvement of speech recognition and natural language process, topic-centric HCIs have achieved great progress and have been widely applied in various applications (cf. (Engwall and Bälter, 2007; Wik and Hjalmarsson, 2009)).Therefore, such agents gained also more attention in the sense of generating corpora based on human-agent communication.

In daily interaction, besides language and speech, humans normally use gestures, gaze, and facial expressions to exchange their intention, contributing to face-to-face communication. Traditional multimodal human-computer dialogue systems were usually constructed according to heuristic lessons from psychology (cf. (Gratch and Marsella, 2005; Rosis et al., 2003; Yang et al., 2014)). These designs help to improve human’s feelings on system interactions, where evaluations were obtained by subjective assessments on task-oriented system feedback analyses (cf. (Bui, 2004; Cassell et al., 1994; Vinayagamoorthy et al., 2006; Yang et al., 2014). Furthermore, aspects like grounding (cf. (Visser et al., 2012)), alignment (cf. (Bergmann et al., 2015)), mimicry (cf. (De Looze et al., 2011)), and understanding (cf. (Traum et al., 2012)) are considered.

For those analyses corpora providing a good quality but also a naturalistic HCI are necessary. Besides interactions with one virtual agent, several datasets are available providing two virtual partners (cf. (Hartholt et al., 2013)). Those are mainly based on the so-called “Utah scenario” (also SOSA4 domain; cf. (Plüss et al., 2011)) providing two agents on one screen. In this paper, based on the multi-agent, multimodal human-computer dialogue system from (Yang et al., 2014), we constructed a corpus adapting the “Utah scenario”. We separated the virtual characters on two screen in slightly different contexts (cf. Section 3.). The whole interaction is in Mandarin.Based on this setup, we are able to consider the already mentioned issues and further, we are able to consider the following two aspects:

Q1: Are there movement patterns identifyable which reflect a certain kind of user’s behaviour?

Q2: Are there any relations between the user’s movement patterns and the current structure of the dialogue or interaction?

2. Dialogue Management

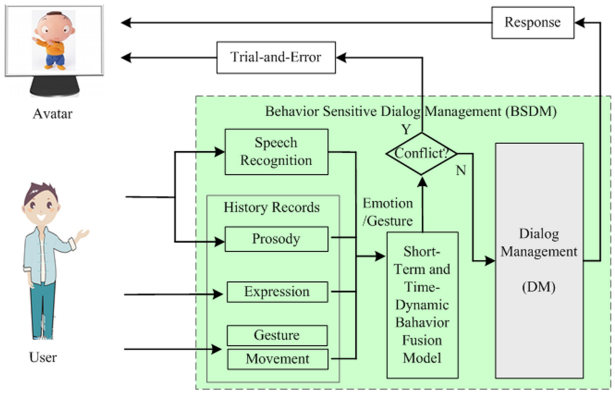

Figure 1: Block diagram of the user behaviour sensitive dialogue management.

The structure of the proposed multimodal human-computer dialogue system is presented in Figure 1, which is similar

to the structure of user behaviour sensitive dialogue management (BSDM) proposed in (Yang et al., 2014). It consists of three components: input module, dialogue management (DM), and output module. Every part in the BSDM receives and deals with multimodal information, including speech, prosody, facial expressions, and gestures.
Multimodal user behaviour features were extracted from a microphone and a camera, which are taken for behaviour fusion in the DM's front-end module with a short-term and time-dynamic (STTD) fusion model (cf. (Yang et al., 2014)). For the user's explicit behaviour, we adopted a keywords list to detect whether the participant's verbal description was in conflict with the other behaviour modalities. Once a conflict was detected the DM asked the user to clarify the intention.
According to the contributions of the different modalities, we distiguished two user behaviours in the DM, namely "explicit behaviour" (obvious interactions and intentions with the system) and "supporting expression" (participant's nonverbal communication and prosodical cues).

3. Recorded Data

3.1. Virtual Coffee House Scenario

Figure 2: The two virtual agents in in the Coffee House scenario interacting with a participant.

Figure 2 presents a virtual coffee house with two virtual agents, where on left the middle-aged male agent is supposed to be its "boss". He is discussing about coffee, drinks, weather, and nearby travelling information to user. The second agent, the so-called "girl", is a female passenger who liked to talk about the weather, introduce further information on the coffee house, and played a game with user. The "boss"' topic flow is listed in the finite state machine (FSM) in Figure 3. The "girl's" FSM is similar.
In our virtual coffee house, equipped with several real furnitures, a Kinect, which is set at the corner between the "boss" and "girl", was used to track user's movements and gestures. An IDS-2CD6024FWD-A/F camera was intended to capture the participant's facial expression, supporting a resolution of more than 200dpi and working well even with poor illumination (cf. Figure 2). To be consistent to the direction of the Kinect, the camera is set between the "boss" and "girl", installed above the Kinect in 170cm from the ground.

3.2. Dialogue Records

The user's behaviour and utterances in the virtual coffee house environment were recorded by a camera, Kinect, and microphone. The recorded items include the participant's gestures, emotional states, and automatically transcribed speech. The response patterns from the "boss" and the "girl" respectively, the chat time and ID (user, "boss", "girl") were logged as well. Dialogue excerpts in Mandarin and English are available on request.
There are currently over 50 users recorded and stored in the corpus, including about 300 effective human-computer dialogue procedures (HCDP). An HCDP means that the user stayed in the virtual house no less than 10 minutes. The average time for each participant in the scenario was about 15 minutes. Totally about 18,000 utterances in Mandarin (including sentences of the user, "boss", and "girl") with corresponding Kinect data are available. This means, that the average number of sentences is about 60 per HCDP.

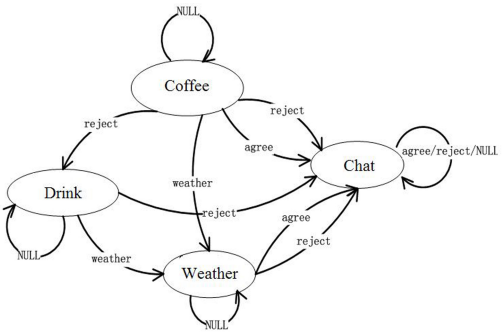

Figure 3: Finite state machine of the topic flow constructed for the male virtual agent called "boss".

4. Preliminary Analyses

To get the analyses of the corpus started, we based the first insights on a randomly selected sub-set of the corpus' material from 10 participants. They interacted with the two virtual agents in a naturalistic way. To enable an automatic processing of the given data, we extracted the relevant information from the log-files provided by the recordings.

4.1. Utilised Material

Based on log-files (cf. Section 3.), we were able to analyse the user's behaviour as well as to show relations between the user's reactions and the current flow of the dialogue. In particular, for this study we used speech input representing the dialogue structure and the Kinect's data reflecting the user's movement and head orientation. In the current experimental setting the log-files provided inputs every second for the measures as well as all utterances spoken by either the participant or an agent. For preliminary analyses this coarse resolution was fair as we were interested in the general interaction's trends and insights on the material.

4.2. Movement Patterns

Since we were interested in the user behaviour during an interaction with the two virtual agents, a part of this behaviour is the participant's movement in the room. This analysis was based on the Kinect data, in particular the 3D values of the spine point provided by the skeleton's structure. Hence, we could approximate the walking activity by plotting the corresponding spine points in the x-z-plane of

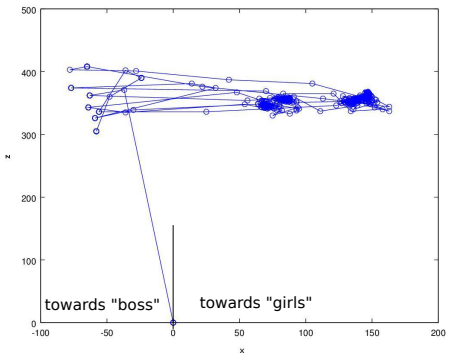

Figure 4: Prototypical plot of the participant's movement in the so-called "band behaviour" group. The sample (0,0) is an artefact resulting from the initialisation of the Kinect.

the Kinect coordinates (cf. Figures 4 and 5), where x indicates the horizontal and z the depth information. The larger the values, the more is the participant oriented towards the female agent. In fact, no temporal context is provided in these plots but the overall activity of the participant. Currently, the movement patterns were analysed manually for the 10 participants in the subset. Later an automatic clustering of the full corpus will be derived. Based on this subset we identified two patterns which reflect a movement behaviour during the interaction.

In the first pattern the participant was fixed to a certain area of the room. The user went only from left to right and vice versa keeping a particular distance from both agents (cf. Figure 4). This pattern will be called "band behaviour". As we could see from the prototype in Figure 4 there are usually two slight attractors were the participant is being fixed. It can be assumed that these small areas are good positions to derive an interaction (cf. Section 4.3.). The "band behaviour" was shown by six participants in the subset.

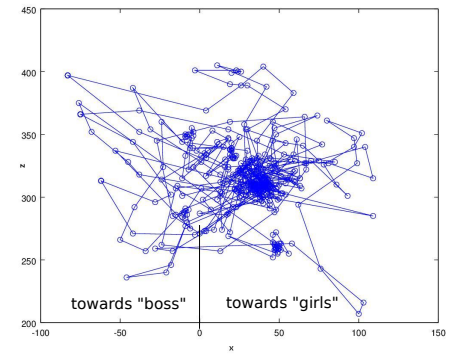

Figure 5: Prototypical plot of the participant's movement in the so-called "chaotic chaotic" group.

In contrast, Figure 5 shows a so-called "chaotic behaviour". The name results from a more differentiated way of walking around in the room using more space in terms of depth. This results in a larger area to be observed during posture and gesture detection. Looking into the single movement plots of the "chaotic behaviour" participants, we found that there is still an area serving as an attractor (cf. Figure 5).

Regarding the two movement patterns, we could conclude that there are some patterns available which are related to a different behaviour of participants. Currently, we identified two characteristics which seem to be reasonable, but further patterns are possible. Our hypothesis is that we will see a small number of prototypical movement patterns in the context of this HCI. From the recent impression, we expect less than 5 patterns in total.

4.3. Combined Analysis

Besides the participant's movement patterns, we also considered the relation between the more generalised user behaviour and the interaction's structure. For this, we combined the movement and head direction information with the temporal dialogue course (cf. Figure 6).

The plots in Figure 6 visualise an example of one participant who is assigned to the "band behaviour". We found that this user is quite stable in terms of x-direction movement while interacting with a certain agent. A further interesting issue was reflected by the head orientation. The participant was still looking at the "boss" even as the "girl" already took the turn. In a more fine-granular temporal resolution as in Figure 6, one could see that the user switched several times between both agents since expecting reactions from the "boss" as well. Finally, a total turn towards the second agent was performed (cf. middle plot in Figure 6). The described phenomena can be seen consistently with all participants.

5. Conclusion

We presented a multimodally recorded, Mandarin corpus where the participants interacted with two virtual agents in a coffee purchasing environment. Currently more than 50 participants were placed in naturalistic communication in a virtual coffee house.

Further, we also presented preliminary results considering the participant's movement patterns while interacting with the agents and provided insights on the dialogue structure. Especially, the later aspects showed that participants tend to be mainly stable in terms of head orientations even if the discussion partner might change, expecting an ongoing communication with the previous interlocutor. This aspect should be considered in the dialogue management and during design of such virtual environments.

Since we currently investigated 10 participants, we will continue the analyses in particular in the dialogue course observations.

6. Acknowledgements

We acknowledge support by the Chinese Academy of Sciences and the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG).

7. Bibliographical References

Bergmann, K., Branigan, H. P., and Kopp, S. (2015). Exploring the alignment space – lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, 2(7). s.p.

Bui, T. (2004). *Creating Emotions and Facial Expressions for Embodied Agents*. Ph.D. thesis, Twente University.

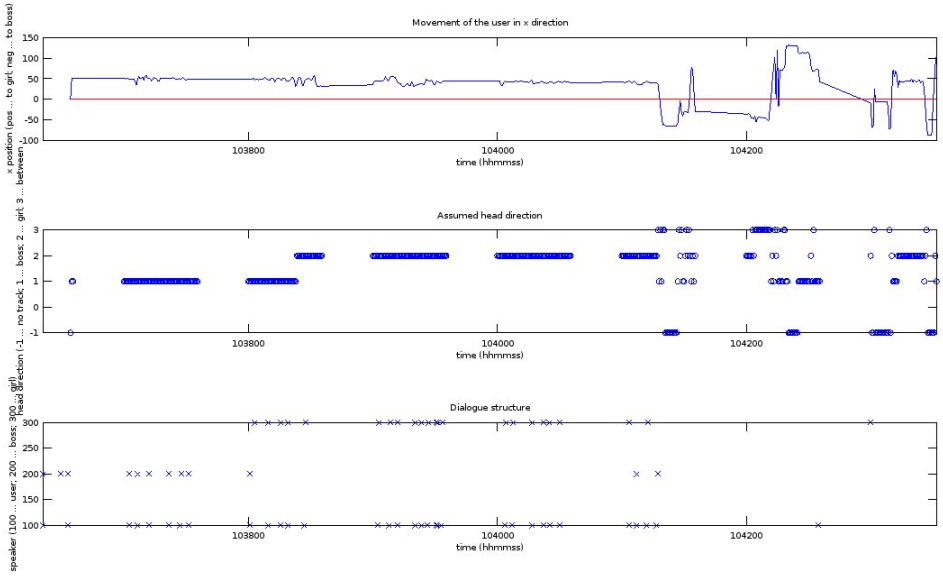

Figure 6: Stacked plot of the participant's movement in x-direction, the head orientation (in the current analysis in combination with the turn of the body), and the dialogue structure in relation to time. For x-direction positive values encode user positions oriented towards the "girl" and remaining values towards the "boss". The head orientation is distinguished between, no track (value -1), to "boss" (1), to "girl" (2), straight to the Kinect (3). In the dialogue structure plot each cross indicates an utterance spoken by the user (100), the "boss" (200), and the "girl" (300).

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proc. of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

Cerekovic, A., Pejsa, T., and Pandzic, I. S. (2009). Realactor: Character animation and multimodal behavior realization system. In Zsófia Ruttkay, et al., editors, *Proc. of the Intelligent Virtual Agents 2009*, volume 5773 of *LNCS*, pages 486–487. Springer.

Courgeon, M., Rébillat, M., Katz, B. F., Clavel, C., and Martin, J.-C. (2010). Life-Sized Audiovisual Spatial Social Scenes with Multiple Characters: MARC & SMART-I^2. In *5èmes Journées de l'AFRV*. s.p.

De Looze, C., Oertel, C., Rauzy, S., and Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *17th Int. Congress of Phonetic Sciences*, Hong Kong, China. s.p.

Engwall, O. and Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learner*, 20(3):235–262.

Gratch, J. and Marsella, S. C. (2005). Lessons from Emotion Psychology for the Design of Lifelike Characters. *Applied Artificial Intelligence Journal*, 19(3-4):215–233.

Hartholt, A., Traum, D., Marsella, S., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., and Gratch, J. (2013). All together now. In Ruth Aylett, et al., editors, *Intelligent Virtual Agents*, volume 8108 of *LNCS*, pages 368–381. Springer.

Kipp, M., Héloir, A., Schröder, M., and Gebhard, P. (2010). Realizing multimodal behavior - closing the gap between behavior planning and embodied agent presentation. In *Proc. of the Intelligent Virtual Agents 2010*, pages 57–63.

Plüss, B., DeVault, D., and Traum, D. (2011). Toward Rapid Development of Multi-Party Virtual Human Negotiation Scenarios. In *Proc. of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011)*, Los Angeles, CA. s.p.

Rosis, F. D., Pelachaud, C., Poggi, I., Carofiglio, V., and Carolis, B. D. (2003). From greta's mind to her face: Modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1-2):81–118.

Traum, D. R., DeVault, D., Lee, J., Wang, Z., and Marsella, S. (2012). Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In *IVA*, volume 7502 of *LNCS*, pages 275–288. Springer.

van Welbergen, H., Reidsma, D., Ruttkay, Z. M., and Zwiers, J. (2010). Elckerlyc - a bml realizer for continuous, multimodal interaction with a virtual human. *Journal on Multimodal User Interfaces*, 3(4):271–284.

Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., and Slater, M. (2006). Building Expression into Virtual Characters. In *Proc. of the Eurographics 2006*. s.p.

Visser, T., Traum, D., DeVault, D., and op den Akker, R. (2012). Toward a Model for Incremental Grounding in Spoken Dialogue Systems. In *Workshop on Real-Time Conversations with Virtual Agents*, Santa Cruz, CA. s.p.

Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication*, 51(10):1024–1037.

Yang, M., Tao, J., Chao, L., Li, H., Zhang, D., Che, H., Gao, T., and Liu, B. (2014). User behavior fusion in dialog management with multi-modal history cues. *Multimedia Tools and Applications*, pages 1–27. online.

A Corpus for a Multimodal Dialog System for Presentation Controls

Paul Hongsuck Seo, Gary Geunbae Lee
Pohang University of Science and Technology (POSTECH)
Pohang, Republic of Korea
{hsseo, gblee}@postech.ac.kr

Abstract

Research studies on multimodal data recently received great interest. Especially combining information in human verbal and gestural modalities is rapidly emerging based on findings of literatures that say the verbal and gestural modalities are highly correlated to each other even in their production. However, there are not many available resources comprising the verbal and gestural modalities making a barrier to start research studies on practical applications. In this ongoing work, we aim to build a multimodal corpus comprising the verbal and gestural modalities designed for a dialog system for presentation controls having eight different features with 18 user intents. The collected data is very rich in modality containing active IR and HD colour videos with audio stream from four microphones array. We present the annotation process of utterances and gestures of the collected presentation recordings for such application with its statistics.

Keywords: multimodality, multimodal corpus, dialog system, presentation control

1. Introduction

Recently, many natural speech interfaces such as Apple Siri, Google Now and Microsoft Cortana, showed their potential of commercial success and thus the field of spoken dialog systems have received great interest and been applied to diverse domains. On the other hand, researches on gesture interfaces have also gotten mature in several applications including games. Microsoft Kinect is a device with sensors for gesture recognition for gaming interfaces. As the studies on these natural user interfaces, speech and gesture interfaces, are getting mature, studies on simultaneous utilization of these two modalities are also receiving large attention since the verbal and gestural modalities often complement each other in human communications. However, there are only few datasets available comprising these modalities (Edlund et al., 2010; Edlund et al., 2012; Oertel et al., 2013).

In this work, we aim to build a multimodal corpus comprising both verbal and gestural modalities. The target domain of the corpus is of a presentation-control dialog system that allows users to control slides by understanding user's co-gesture speeches during a presentation. Because studies are lacking on processing gestures in general taxonomy where gestures are mapped into generic semantic phrases and linking to linguistic phrases, we instead targeted a specific domain where the semantic meaning of gesture phrases can be extracted more easily and clearly.

2. Co-gesture Speeches

Co-gesture speeches are speeches accompanying natural gestures often for delivering messages more clearly and play the key roles as the cardinal elements in the cross-section of both modalities. While manual and head gestures are studied most and are mostly synchronized to spoken signals semantically and pragmatically, manual gestures also have a capacity to add supplementary contents that are often more effective than speeches (Wagner et al., 2014). For instance, saying "bring that" while pointing an object can allow the speaker to avoid from making precise verbal description for the location of target objects. Many literatures argue that gestures and speeches are strongly correlated to each other even in their production (Wagner et al., 2014).

	Natural co-gesture speech triggers		Guided gesture triggers
	speeches	gestures	
Page Navigation	Primary	Not used	Defined
Video Control	Primary	Secondary	Not defined
Pointer	Not used	Primary	Not defined
Zooming	Primary	Secondary	Defined
Hyperlink	Primary	Secondary	Not defined
Object Control	Primary	Secondary	Not defined
Annotation	Primary	Primary	Not defined

Table 1: List of final design features with modalities used

Since we mainly focus on building a corpus for a dialog system where understanding user's intents is important, speeches and gestures are semantically mapped through the intents.

3. Presentation Control Dialog System

The target application of this work is a dialog system that controls the presentation slides according to user's co-gesture speeches. We conducted a participatory design, in which end-users actively join the design process, for such application with five participants who frequently make presentations (twice a month in average). The final design consists of seven different features: page navigation, video control, pointer, zooming, hyperlink, object control and annotation features.

Gestures for features can be divided into two types: primary guided gestures, primary natural gestures and supplementary natural gestures. Since gestures are not commonly the main modality for human communications,

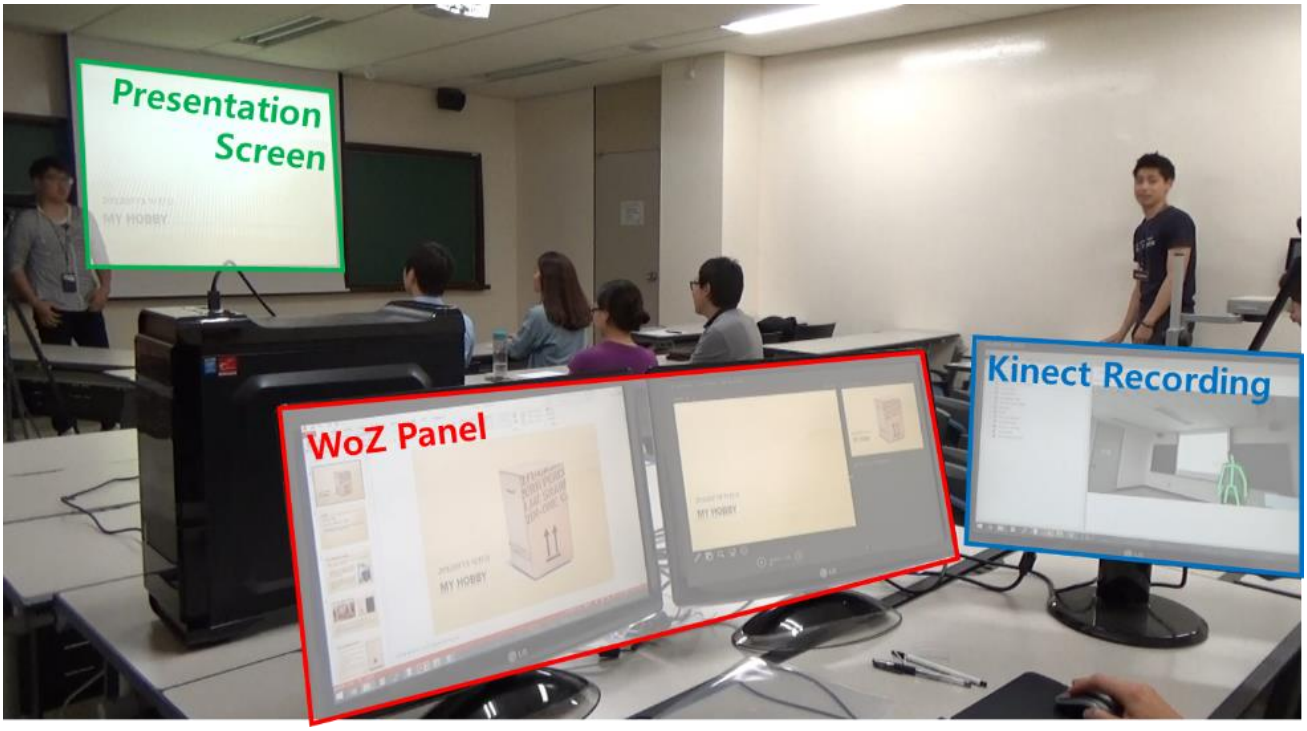

Figure 1: WoZ setting with Kinect recording

natural gestures mostly play secondary roles supplementing verbal modality in spatial dimension. Pointing gesture is the only exception that a natural gesture is used as the main modality. Pointing gestures could be used as either primary or secondary modality. When the user points the screen explaining the topic of presentation, it is natural primary expression of directing the related spot on the screen and at the same time expressing a user intent for the system to react by showing a pointer on the spot. On the other hand, when the user points a video on a slide and says "play this video", pointing gesture is used to complement the main modality, the verbal modality, through which the user expressed the intent. In addition to the natural co-gesture speeches, we defined guided gestures that are predefined, command-like gestures for triggering a system action. In this case, gestures play the primary role for expressing intents. These guided gestures are designed because expressing the intents for every control is unnatural even with natural co-gesture speeches. The list of the final design features is shown in Table 1 with the information of modality use for each feature.

4. Data Collection

To reduce the difference between user's behaviours in the collected dataset and the real inputs from the deployed system, we collected the dataset under the Wizard of Oz (WoZ) setting. The recruited subjects were asked to prepare a five-minutes-long presentation of delivering information they know more than other people do. Before their making a presentation, they were informed of the features that they can use to control their slides during the presentation. In total, 12 different subjects made 16 different presentations. The average duration of the presentations was about ten minutes including Q&A sessions and every presentation was in Korean.

We recorded all the presentations using a Microsoft Kinect v2.0 sensor which has an active infrared (IR) sensor (512×424; 30Hz; depth to 8m), a HD colour sensor (1920×1080; 30Hz or 15HZ in low light) and a microphone array (4 microphones). We used Microsoft Kinect Studio v2.0 for recording all the sensed input sources in raw format. Hence, the collected dataset contains very rich information in both modalities. In visual modality, mapping inputs from IR and colour sensors gives us 3D spatial information. In auditory modality, noise cancelled signals with the angle to the sound source can be obtained by beamforming the inputs from the microphones of the array. Since every sensor outputs are stored in their raw format, the size of the collected dataset is about 1.46TB, which is quiet large. The WoZ setting with Kinect recording is shown in Figure 1.

5. Annotation

In this section, we describe the annotation process on the collected dataset and the statistics of the annotation result.

5.1 Annotation Scheme

We used an XML format. Figure 2 is a part of our annotated corpus. The verbal modality and the gestural modality are annotated in separate tags: <*utters*> and <*gestures*>. An <*utters*> tag contains all of the transcribed speaker's utterances sentence by sentence each tagged by an <*utter*> tag. Each utterance tagged by an <*utter*> tag is aligned with its corresponding speech input from the microphones through the attributes *start_time* and *end_time*.

Similarly, a <*gestures*> tag contains <*gesture*> tags that align speaker's gestures to their corresponding visual signals. As in <*utter*> tags, <*gesture*> tags also have the

```
<presentation ks_file='KS_p001'>↓
    <utters>↓
        <utter start_time='0:00:08.2' end_time='0:00:09.2' intent='NONE'>↓
            네 안녕하세요.↓
        </utter>↓
        <utter start_time='0:00:09.2' end_time='0:00:10.6' intent='NONE'>↓
            저는 P2입니다.↓
        </utter>↓
        <utter start_time='0:00:11.5' end_time='0:00:14.3' intent='NONE'>↓
            어, 네. 오늘 제가 소개할 취미는요.↓
        </utter>↓
        <utter start_time='0:00:15.6' end_time='0:00:15.7' intent='NONE'>↓
            이겁니다.↓
        </utter>↓
    </utters>↓
    <gestures>↓
        <gesture start_time='0:00:14.2' end_time='0:00:15.0' type='swipe' intent='PN-next' />↓
    </gestures>↓
</presentation>↓
```

Figure 2: Example of XML annotation

same attributes *start_time* and *end_time* for the alignment. While the contents of <*utter*> tags are sentences, which can be a reasonable and human-understandable meta-representation of the actual audio signal, we used the *type* attribute for the meta-representation of the actual visual signals. There are five possible values for the *type* attribute chosen based on the target domain and the design result of the application:

Pointing: gestures of pointing the screen

Swipe: gestures of moving a hand from left to right or right to left in front of the speaker's body

Pinch: gestures of pinching a hand directing a spot of the screen

Stretch: gestures of stretching a hand directing a spot of the screen

Scroll: gestures of moving a hand at the screen from side to side

This means that only these gestures corresponding to the above five possible types are annotated. Note that all of the target gestures are manual gestures because of two reasons: Firstly, the natural co-gesture speeches used by the speakers during the data collection only contained pointing gestures. Secondly, the guided gestures predefined by the system designers were all manual gestures.

Finally, the <*utter*> and <*gesture*> tags have an important attribute *intent* representing the corresponding user intention for controlling slides. There are in total 18 possible intents with *NONE*. The code *NONE* stands for the case of having no specific intent in the target domain and is used only for <*utter*> tags but not for <*gesture*> tags. It is because we annotated only five types of gestures, which explicitly contain domain specific meanings. However, in the case of user utterances, all utterances said by the speakers are transcribed so there are more utterances for the presentation itself than ones for the controls of slides. Those transcribed out-of-domain utterances are annotated with a *NONE* tag.

Page Navigation	
PN-prev	Move to previous slide
PN-next	Move to next slide
PN-rand	Move to specified target slide
Video Control	
VC-play	Play video clip
VC-stop	Stop playing video clip
VC-prev	Move to previous bookmark of video
VC-next	Move to next bookmark of video
VC-rand	Move to specified target bookmark
Pointer	
PT-point	Show pointer on screen
Zooming	
ZM-in	Zoom in area on screen
ZM-out	Zoom out to whole slide
ZM-move	Move zoomed area
Hyperlink	
HL-open	Open hyperlink
HL-close	Close hyperlink
HL-scroll	Scroll on hyperlink page
Object Control	
OC-move	Move target object to specified place
OC-large	Enlarge object
OC-small	Shrink object in size

Table 2: List of user intents for presentation controls

5.2 Annotation Results

The total number of presenter's utterances in all the recordings is 1,086 and the average number of words per a utterance is 10.2. Among these utterances, only 126 utterances contained in-domain user intents and seven of them are annotated with two intents because with these utterances, presenters expressed multiple intentions in a single utterance. The other 960 utterances are purely for the presentations themselves and are annotated with a *NONE* tag. On the other hand, only in-domain gestures are annotated and their total number is 629. The number of annotations for each intent is shown in Table 3. While the presenters mostly used a single modality for expressing intents, 42 intents (29 PT-point, 8 ZM-in, 3 HL-open and 2 HL-scroll) are expressed through both modalities together complementing each other modality.

Interestingly, we could observe that the distributions of intents for utterances and gestures greatly differ from each other (Figure 3). This difference evidently show that two modalities are better utilized for different purposes. A half of the gesture usage is for pointing. This follows the argument that the gestural modality is often better for expressing spatial information (Wagner et al., 2014). A large portion of the other half is for page navigation to the next page, which is a guided manual gesture.

	Utterances	Gestures
NONE	960	0
PN-prev	5	8
PN-next	10	263
PN-rand	24	0
VC-play	21	0
VC-stop	4	0
VC-prev	0	0
VC-next	2	0
VC-rand	10	0
PT-point	30	306
ZM-in	10	26
ZM-out	5	19
ZM-move	0	1
HL-open	7	3
HL-close	3	0
HL-scroll	2	3
OC-move	0	0
OC-large	0	0
OC-small	0	0
Total	1086	629

Table 3: Numbers of intents annotated to utterances and

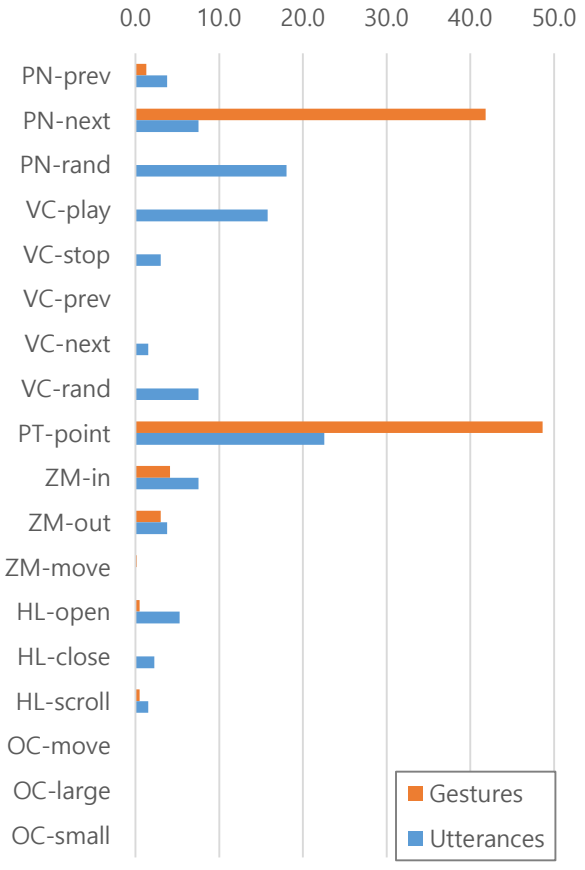

Figure 3: Distributions of user intents annotated to utterances (blue) and gestures (orange). Out-of-domain utterances are excluded in the distribution.

6. Discussion

In this work, we focused on collecting recordings comprising multimodal input sources and annotating user intents to implement a multimodal dialog system. However, this corpus should still be extended to contain the information of the target object, i.e. slot values in the dialog system. This is not an easy task differing from previous dialog corpora since the target is an object on the screen, which are weakly or never expressed verbally. In previous dialog corpora, the aim was to find the words referring the slot values. In this work, the slot values are more often designated by gestures than by verbal expressions.

Although the corpus was designed for a specific application, a multimodal dialog system for presentation controls, the corpus draws several other research interests such as investigating relation between human speech and its natural co-gestures, and improving automatic speech recognition using multimodal inputs.

7. Conclusion

We collected a multimodal corpus for a dialog system for presentation controls mainly focusing on co-gesture speech interactions. We collected recordings of 16 presentations made by 12 different speakers in WoZ setting using Microsoft Kinect v2.0 for rich-multimodality. After collecting the recordings, we transcribed all 1,086 utterances of the presenters and annotated 629 domain specific gestures. Then, each utterance or gesture are tagged with its corresponding user intent for the designed features of a presentation tool.

8. Acknowledgement

This research was supported by the ICT R&D program of MSIP/IITP. [B0101-15-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)]

9. References

Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson and David House. Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture. *In Proceedings of LREC*; 2010. p. 2992-2995.

Jens Edlund, Mattias Heldner and Joakim Gustafson. Who Am I Speaking At? Perceiving the Head Orientation of Speakers from Acoustic Cues Alone. *In Proceedings of*: LREC; 2012.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner and Nick Campbell. D64: A Corpus of Richly Recorded Conversational Interaction. *Journal on Multimodal User Interfaces*. 2013;7(1-2):19-28.

Petra Wagner, Zofia Malisz and Stefan Kopp. Gesture and Speech in Interaction: An Overview. *Speech Communication*. 2014;57:209-232.

Integrated Health and Fitness (iGF)-Corpus - ten-Modal Highly Synchronized Subject-Dispositional and Emotional Human Machine Interactions

Michael Tornow[1], Martin Krippl[2], Svea Bade[1], Angelina Thiers[3], Ingo Siegert[1], Sebastian Handrich[1], Julia Krüger[4], Lutz Schega[3], Andreas Wendemuth[1]

[1]Institute of Information and Communication Engineering, [2]Department of Methodology, Psychodiagnostics and Evaluations Research, [3]Institute of Sport Science, [4]Department of Psychosomatic Medicine and Psychotherapy, Otto von Guericke University, 39016 Magdeburg, Germany

Abstract

A multimodal corpus on human machine interaction in the area of health and fitness is introduced in this paper. It shows the interaction of users with a gait training system. The subjects pace through a training course four times. In the intermissions, they interact with a multimodal platform, where they are given feedback, they re-assess their performance and they plan the next steps. A high involvement of the subjects is given. By design, the interaction further evokes cognitive underload and overload and emotional reactions. The platform interaction was arranged as a Wizard of Oz Setup. In the interaction phase, 10 modalities are recorded in 20 sensory channels with high performance of hardware synchronicity, including several high-resolution cameras, headset and directional microphones, biophysiology, 3D data as well as skeleton and face detection information. In the corpus, 65 subjects are recorded in the interaction sessions for a total of 100 minutes per subject, including self-ratings from eight time points during the experiment. Additionally, several questionnaires are available from all subjects, regarding personality traits, including technical and stress coping behavior.

Keywords: Corpus, Interest, Involvement, Cognitive Overload, Cognitive Underload, Human-Machine-Interaction, Multimodality, Health, Fitness, Planning

1. Introduction

The area of integrated Health and Fitness is gaining increased attention both in popular lifestyle as well as for prevention of prospective diseases, in particular for persons in their midlife or elderly life who show no medical diagnosis but may belong to a risk group. In this field, we present the iGF-corpus (German: *integrierte Gesundheit und Fitness*) showing the interaction of subjects with a gait training system, as gait is known to be a reliable lead indicator for posture-related later diseases (Hamacher et al., 2015).

The corpus is a rich source of study for general feedback, planning and interaction activities in multiple modalities in Human Machine Interaction (HMI). Its main assets are elaborated hardware synchronicity over 10 modalities recorded in 20 sensory channels, and dedicated and standardized phases of subject-dispositional reactions (interest, cognitive underload andcognitive overload) as well as standardized HMI-related emotional reactions (fear, frustration, joy) with the same group of elderly subjects.

In HMI, interaction modalities can be used as implicit or explicit inputs. Explicit inputs are used for direct control, whereas implicit inputs supply the machine with information about the emotional state or the disposition of the subject, which can favorably be used for subject adaptation and which serve as an important component for companion technology (Biundo and Wendemuth, 2016). These states can be detected in a wide range of signals emitted by the human body, e.g. prosody, body posture, gestures, mimic or biophysical signals.

The scenario is given by subjects pacing through a training course four times. In the intermissions, they interact with a multi-modal platform, arranged as a Wizard of Oz Setup (Kelley, 1984). 65 subjects were recorded in the interaction for a total of 100min each. Personality information is obtained from selfratings and questionnaires.

2. Theoretical Background

The aim of the present corpus is to provide multi-modal data of subjects expressing selected dispositional states (interest, cognitive underload and cognitive overload) as well as HMI-related emotional reactions (fear, frustration, joy).

Interest is understood as a multi dimensional construct of cognitive concern or attention of a person for an object or a person (Deci and Ryan, 1985). The degree of interest is defined as degree of appreciation of the activities. As it is still unclear how to elicit interest, we rely on the sub-construct "notice". This term describes the perception and selection of certain environmental stimuli. As human stimuli processing is restricted, only certain information can achieve consciousness. Recent research leads to the insight that selective interest can be controlled by appealing topics and motivation (Ainley et al., 2002; Giakoumis et al., 2011).

Cognitive Underload is mostly connected to boredom, where a subject is willing, but being unable, to engage in satisfying activity (Eastwood et al., 2012). To generate such a state, the activity must provide non-engaging and non-satisfying activities covering passivity or monotony. The key is a low peculiarity of novelty of the interaction situation (Giakoumis et al., 2011). A challenge for inducing this state is to prevent subjects from experiencing anger, impatience or tension due to their motivation to fulfill the task.

Cognitive Overload describes states where the working memory capacity has reached its limit (Miller, 1956). A cognitive overload increases the risk of error in the actual task (Chandler and Sweller, 1992). To achieve such a situation, subjects must gain several stimuli while working on complex tasks. Especially for elderly subjects heavy cognitive load is reached earlier, as aging contributes to a decline in the efficiency of working memory.

Standardized Emotion Induction The whole experiment is concluded by an emotion induction using standardized

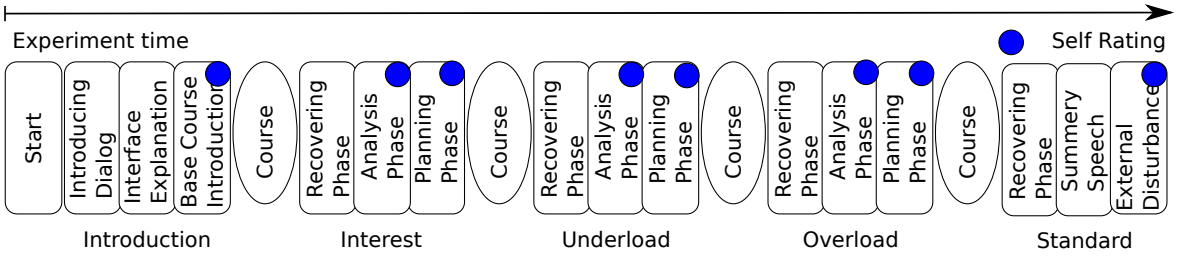

Figure 1: The overall procedure of the experiment

methods to gain information on the expressiveness of the subjects. For this purpose, we selected HMI-relevant emotions: fear (Brave and Nass, 2003), frustration (Why and Johnston, 2008), and joy (Brave and Nass, 2003).

3. Experiment Design and Setup

One major problem in recording a dataset on natural mental states is often the missing involvement of the subjects. We resolved this issue by including the data recording into a health and fitness scenario. The actual system is introduced to the subject as a training system for human gait which should be optimized with the help of the subjects. The usage as well as the exercises and the subject task are explained by the system itself.

Subjects were presented an automatic gait analysis system, composed of two parts, a gait training course (cf. Fig. 2 left) and a technical interface for planning and evaluating the gait training. The training course was physically constructed and all subjects had several runs through the changing course. The data of the gait training is not part of the corpus. The interaction with the technical system took place in a separate area and is the basis for this corpus. The target group are subjects with the age of 50 and above but without any known problems in their gait.

The course of the experiment is divided in five modules each oriented on a special topic (cf. Fig. 1). After the experiment start and introduction module [1], the subjects undergo the three cognitive elicitation modules interest [2], underload [3], and overload [4]. The experiment is completed by the emotion induction module [5]. The whole interaction is standardized and all subjects underwent the interaction in the same order.

To retain the subject's involvement the gait course is placed after each of the modules 1-4. To avoid an interference in the biophysical measurements of emotional induction and gait exercise it is followed by 5min relaxing phase. Subjects listened to Vivaldi's *The Four Seasons* as Thompson et al. (2005) show that it has an relaxing effect. Furthermore, the modules 2 to 4 used (simulated) analysis of the subject's gait related to his last gait course, followed by a planning phase in which the subject plans a new gait course.

During the Wizard of Oz experiments the subject uses voice commands to interact with the system that is able to answer using a Text-to-Speech system and a graphical interface providing additional information (cf. Fig. 2 center). Furthermore, the subjects are recurrently asked to rate their own emotional state along valence, arousal, and dominance dimensions using the Self Assessment Manikins (Bradley and Lang, 1994). The self rating is placed on a distinct tablet computer, to reduce the influence on the experiment. In the following, we shortly describe the purpose and implementation of the different modules:

Module 1: Introduction In this module the gait training system and its usage are introduced. The purpose of this module is to make a subject familiar with the way of communicating with the system for getting a more natural and expressive behavior. Therefore a set of questions focused on personal details: age, body size, place of residence, profession, family as well as technical affection are asked. Furthermore, the first gait exercise device is introduced. This phase is closed by a self-rating.

Module 2: Interest This module starts with a relaxing phase followed by an analysis phase and a planning phase. In the analysis phase the subject receives information on 12 newly introduced gait exercise devices. Four information categories are available (general information, exercise instructions, physical load/error patterns, relation to every day life). In the planning phase six of the thirteen training elements have to be used to create an individual gait course. After course planning, the subject is again asked for a self-rating. Meanwhile a team of assistants prepared the planned course.

Module 3: Cognitive Underload After the relaxing phase, a reduced interaction in comparison to module 2 is used for the analysis phase and the system gave only a very simplified analysis. Afterwards, the system confronts the subject with a learning situation, where very similar sentences providing barely varying information on the thirteen exercise devices should be read several times. This very boring, but important task induced cognitive underload. The planning phase is exactly like the one in the interest module.

Module 4: Cognitive Overload At the beginning, the subject had a relaxing phase of 5 minutes. During analysis, negative feedback is provided regarding the subject's gait during the last course. Furthermore, the system expresses dissatisfaction with the subject's analysis of the training progress. Overload is induced by asking for information on a topic with which the subject is not very familiar with. In the planning phase, the subject should create a course avoiding hip overstraining. This constraint results in exactly one combination of exercise elements feasible, which the subject is not aware of. Additionally a time constraint solving the planning in 45 seconds is introduced. These two constraints result in a very stressful situation.

Module 5: Standardized Emotion Induction The last module starts again with the relaxing phase. But instead of

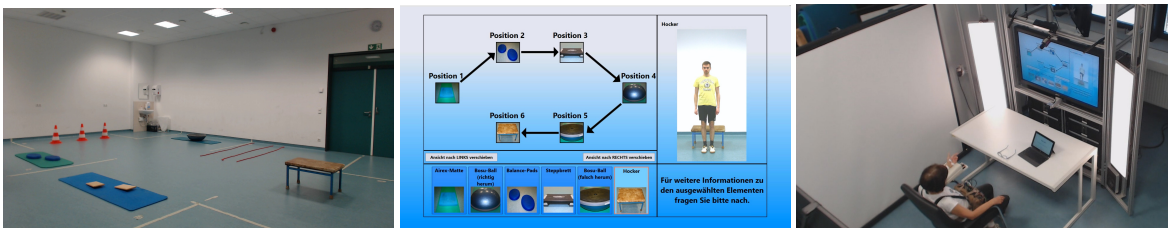

Figure 2: Left: Gait Training Course Example; Middle: The Screen of the Course Planning; Right: Sensorial Setup.

the analysis and planning phase, the subject is faced with the task to summarize his impressions and learned information using a short talk of about two minutes. To increase the stress level, a subsequent examination by the system was announced. Too short talks are not accepted by the system. Fear is elicited by a sudden alarm sound from a device somewhere in the training room.

4. Questionnaires

Before and after the experiment, several measures of personality and mood were applied: **TAT/PSE** to measure the implicit basic needs of the participants (McClelland et al., 1989), **PRF** to measure dominance, affiliation and achievement (Stumpf et al., 1985), **BMPN and Autonomy-Scales** to measure the need to be autonomous (Deci and Ryan, 1985; Krippl, 2015), **TA-EG** to measure the technology affinity (Karrer et al., 2009), **PANAS** to measure the subject's actual mood (Kuhl and Fuhrmann, 1998).
HAKEMP was used to measure action vs. state orientation (Kuhl, 1990), **NEO-FFI** to measure the 'big five' of personality (Borkenau and Ostendorf, 1993), **SSI-K** to detect skills and deficits in the field of self-control and the will for physical training (Kuhl and Fuhrmann, 1998), **ERQ** to measure the emotion regulation abilities (Abler and Kessler, 2009). Furthermore, a self constructed measure for the occurence of walking problems was used.

5. Technical Setup

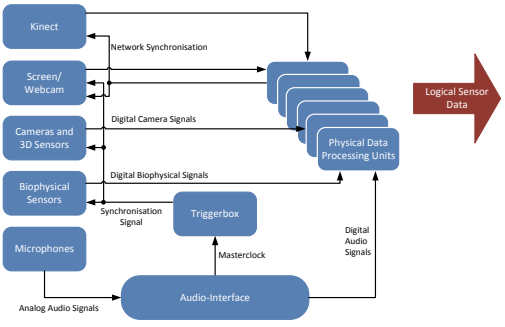

Figure 3: Data Flow and Synchronization.

The entire subject interaction is recorded with a synchronized multimodal sensor phalanx set up as a network controlled distributed system (cf. Fig. 3). To guarantee the synchronicity a trigger signal of 25Hz is generated from the audio sampling rate of 44.1kHz of the audio interface Yamaha 01V96i by dividing it by 1764 using a trigger box consisting of the USB-controlled timer counter device Measurement Computing USB4303 and some amplifiers for signal adoption. This trigger signal is initialized from a control center at the start and shut down at the end of the recording and directly used to control the camera recording. It contains three AVT Pike P45C cameras (upper, lower and left camera) equipped with a 16mm lens.

Modality	Parameter	Alignment
Video		
Pike	3xRGB 1388x1039px, 25Hz	22.6μs
Webcam	RGB 1920x1080px, 30Hz	22.6μs
Screen	RGB 1920x1080px, 30Hz	22.6μs
Kinect2	RGB 1920x1080px, 30Hz	70 ms
Body-Posture		
Posture	25 body points, 30Hz	70 ms
Face	5 Points, 30Hz	70 ms
Kinect2	3D-Data 512x424px, 30Hz	70ms
Kinect2	IR 512x424, 30Hz	70ms
Audio		
Proband	4xMono, 16Bit, 44.1kHz	22.6μs
Wizard	1xStereo, 32Bit, 44.1kHz	22.6μs
Kinect2	1xStereo, 32Bit, 44.1kHz	70ms
Biophysiology		
EMG	16Bit, 256Hz	4ms
ECG	16Bit, 256Hz	4ms
EDA	16Bit, 256Hz	4ms
Annotation		
Marker	8Bit, 44.1kHz	22.6μs

Table 1: Modalities and their parameters

To gain an overview on the subject's actions the scene is recorded using a webcam under the ceiling of the training room as well as the subject interaction screen. Along with those the trigger signal is recorded in an audio track allowing to correct the alignment of the modalities in a post-processing step. The recording of the Microsoft Kinect2, providing a a number of modalities is started and synchronized via network control according to the webcam and the screen recording with an accuracy of 70ms. The auditive output is recorded in Cubase via the audio interface on 6 tracks, containing the signals of a radio headset Sennheiser HSP2 via EW 100, two shot gun microphones Sennheiser ME66 and the left and right channel of the interface output. By recording the subject's and the interface's auditive signals separately we can identify cross talk and dominance in the interaction. Furthermore, a marker channel is recorded where the timeline of events and the Wizard text can be reconstructed. The psychophysical data streams ECG, EMG and EDA are recorded using the mobile biophysical mea-

surement system gtec Mobilab II+ connected via Bluetooth using a Laptop and the software recorder (cf. Table 1).

6. Dataset Characteristics

Subjects/Experiments	65 (41 with all modalities)
Gender	Male 20 / Female 45
Total Recorded Data	105 h 48 min (66 h 43 min)
Experiment Duration	Mean: 97 min
Age	over 50 (Min: 50; Max: 80; Mean: 66)
Language	German
Annotation	Events, System's Speech Output

Table 2: Dataset Characteristics

Table 2 summarizes the dataset characteristics. Researchers interested in the corpus please contact *igfcorpus@ovgu.de*. For non-profit scientific work the corpus is provided without charge according to the terms of use.

7. Conclusion

In this paper a new dataset on natural HMI is proposed, whereas a high involvement of the elderly subjects is realized using a gait training system. Within the course of the experiment the subjects face interesting situations as well as cognitive underload and overload. Ten modalities within the interaction are recorded in 20 synchronized sensory channels controlled by a network based distributed recording system. The data includes multiple high resolution video and audio data streams, as well as 3D and infrared information recorded by a Kinect2. Furthermore, the biophysical data channels ECG, EMG and EDA are recorded. In total, 65 subjects take part in the experiment whereas the recording of 41 is complete over all modalities and induction modules. The mean recording time per person is about 100 Minutes giving 105 hours of recorded material. The dataset will be enriched with additional modalities gained from post-processings.

8. Acknowledgements

The work presented was done within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG). The authors thank Michael Kotzyba, Stefanie Rukovina, Markus Kächele, Matthias Haase, Sascha Meudt, Stefanie Rukavina for the support during the experiment design. We thank Günther Palm, Michael Glodek, Jörg Frommer, and Michael Weber for supervision. The technical background crew Ralph Heinemann, Jens Holze and Sebastian Stroutz was invaluable for setting up and maintaining the recording set up. Numerous research assistants supported the realization.

9. References

Abler, B. and Kessler, H. (2009). Emotion regulation questionnaire-Eine deutschsprachige Fassung des ERQ von Gross und John. *Diagnostica*, 55(3):144–152.

Ainley, M., Hidi, S., and Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *J Educ Psychol*, 94(3):545.

Biundo, S. and Wendemuth, A. (2016). *Companion Technology - A Paradigm Shift in Human-Technology Interaction*. Springer International, in press.

Borkenau, P. and Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae-Deutsche Fassung*. Göttingen: Hogrefe.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J Behav Ther Exp Psy*, 25(1):49–59.

Brave, S. and Nass, C. (2003). Emotion in human–computer interaction. In *Human-Computer Interaction*, pages 81–96.

Chandler, P. and Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62(2):233–246.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self- determination in human behavior.*

Eastwood, J. D., Frischen, A., Fenske, M. J., and Smilek, D. (2012). The unengaged mind defining boredom in terms of attention. *Perspectives on Psychological Science*, 7(5):482–495.

Giakoumis, D., Tzovaras, D., Moustakas, K., and Hassapis, G. (2011). Automatic recognition of boredom in video games using novel biosignal moment-based features. *IEEE Trans. on Affect. Comp.*, 2(3):119–133.

Hamacher, D., Hamacher, D., Singh, N. B., Taylor, W. R., and Schega, L. (2015). Towards the assessment of local dynamic stability of level-grounded walking in an older population. *Med Eng Phys*, 37(12):1152 – 1155.

Karrer, K., Glaser, C., Clemens, C., and Bruder, C. (2009). Technikaffinität erfassen – der Fragebogen TA-EG. *Der Mensch im Mittelpunkt technischer Systeme*, 8:196–201.

Krippl, M. (2015). *Die Autonomiebedrüfnis-Skalen*. Universität Magdeburg.

Kuhl, J. and Fuhrmann, A. (1998). *Das Selbststeuerungsinventar (SSI)*. Universität Osnabrück.

Kuhl, J. (1990). Der Fragebogen zur Erfassung von Handlungs-versus Lageorientierung (HAKEMP 90). *Unveröffentlichter Fragebogen, Universität Osnabrück*.

McClelland, D., Koestner, R., and Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychol Rev*, 96:690–702.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D., and Beloch-Till, H. (1985). *Deutsche Personality Research Form (PRF)*. Göttingen: Hogrefe.

Thompson, R., Moulin, C., Hayre, S., and Jones, R. (2005). Music enhances category fluency in healthy older adults and Alzheimer's disease patients. *Exp Aging Res*, 31(1):91–99.

Why, Y. P. and Johnston, D. W. (2008). Cynicism, anger and cardiovascular reactivity during anger recall and human–computer interaction. *Int J Psychophysiol*, 68(3):219–227.

A Robotic Exploration Corpus for Scene Summarization and Image-Based Question Answering

Claire Bonial, Taylor Cassidy, Susan G. Hill, Judith Klavans, Matthew Marge, Douglas Summers-Stay, Garrett Warnell, Clare Voss
U.S. Army Research Laboratory
2800 Powder Mill Rd, Adelphi, MD 20783
Claire.N.Bonial.civ@mail.mil

Abstract

The focus of this research is the development of training corpora that will facilitate multimodal human-robot communication. The corpora consist of video and images recorded during a robot's exploration of an office building and several types of associated natural language text, gathered in four distinct annotation tasks. Each annotation task supports the development of training data for a foundational technology integral to facilitating multimodal communication with robots, including object recognition, scene summarization, image retrieval, and image-based question answering. Although each of these technology areas has been addressed in work on computer vision, our research examines the unique requirements of these technologies when the visual data is collected by a robot and analyzed from a first-person perspective. For our purposes, the robot must be able to provide efficient natural language summaries of what it is seeing and respond to natural language queries with visual data. Here, we describe the progress of this ongoing research thus far, including the robotic exploration data and the development of annotation tasks.

Keywords: scene summarization, question-answering, human-robot communication

1. Introduction

This research focuses on the challenges of associating natural language with video and image data collected from a mobile robotic system. Today, robots offer unique capabilities as teammates in a variety of tasks, such as search and rescue operations. One value of robots is that they can move into environments that are inaccessible to people for any variety of reasons. Taking advantage of this potential to its fullest, however, requires efficient communication between robots and remote humans. Given the limited communication channels available in the midst of natural disasters or war-torn environments, efficient communication must be flexible in what form it takes, allowing for multimodal communication. Here, we focus on two modes of communication: natural language and images. Natural language has the benefit of familiarity and flexibility for people, but in some contexts, the information contained in an image is more informative and reliable than a natural language description.

For robots to be more effective than mobile security cameras, it is imperative that robots summarize overwhelming streams of data in a way that humans can quickly understand, facilitating decision-making tasks. Thus, one goal of this work is to develop automated techniques for generating natural language summaries of a robot's first-person visual information. In addition, we need the ability to query the robot regarding what it has seen in its autonomous exploration of unknown environments and receive relevant images in response to those natural language queries. Thus, another primary goal of this work is to model the likelihood that a given image provides enough information to answer a natural language question.

Our approach to these problems begins with the collection of video and image data by teleoperating a mobile robot equipped with a camera and laser sensor through an office environment. The video and images are associated with a variety of natural language annotations obtained through crowdsourcing on Amazon's Mechanical Turk.[1] This research is currently underway; here we describe initial data collection efforts and the development of four distinct annotation tasks needed to construct the training corpora. We then compare this research to related work, and close with a summary of contributions.

2. Corpora

This work will culminate in four distinct corpora (or "datasets"). Each corpus consists of visual data and natural language text. However, each corpus is designed to serve as training data for one component in what will eventually be a multimodal human-robot communications system. Our goal is to expand the capabilities found in existing human-robot dialogue frameworks (e.g., Lemon et al., 2001) that can process natural language input and execute commands. Our two primary objectives are for a robot to be able to (i) efficiently summarize what it is seeing in natural language, and (ii) retrieve images that enable a human to answer a natural language question. The four annotation tasks and datasets are summarized in Table 1, and both the visual data and annotation tasks are described in detail in the sections to follow. We plan to release these multimodal corpora to the research community.

2.1. Robotic Exploration Data

The video and image data used in this research will be collected by a custom build of the Clearpath Robotics Jackal robot (see Figure 1) that is currently equipped with an RGB camera capable of recording VGA-quality video and a high-resolution 2D laser scanner. The robot uses laser

[1] http://www.mturk.com

Dataset	Annotation Task Description	Component Technology Supported
1	Object part, feature and property labeling	Object Recognition
2	Summaries of sequences of images, static scenes	Scene Summarization
3	Elicitation of questions from Turkers	Image-Based Question Answering
4	Selection of adequate/ideal images to answer a question	Image Retrieval, Image-Based QA

Table 1: Annotation tasks and datasets being collected for robotic exploration.

Figure 1: The custom-built Jackal robot used in video/image data collection.

scanner data to create a 3D environment map using available simultaneous localization and mapping (SLAM) software, the output of which will be included in our dataset.

The Jackal's dimensions are 20 x 17 x 10 inches. Software on-board the Jackal limits its maximum speed to 1.5 meters per second (approximately 3 miles per hour). Researchers teleoperate the Jackal using an off-the-shelf game controller.

As part of our research supporting scene summarization, we will take as input a stream of image sequences, the time-aligned positional information, and SLAM data. A process will automatically determine the most informative (i.e. high entropy) images in a sequence, which will be used to produce initial text captions. We will combine multiple captions into a broader summary using natural language generation techniques. For a given scene and a given natural language question pertaining to that scene, each image from the robot's video stream will be manually labeled as either answering the question or not. Each image will be associated with the relative location and orientation (pose) of objects of interest and the robot, which are automatically derived using the robot's 3D map.

So far, we have collected video and image data during exploration of a library within an office building. For the environments explored, AprilTags (Olson, 2011), a type of fiducial marker, were placed in predefined canonical locations on desks, books, plants, cabinets, a bookshelf and a trophy case for the purpose of facilitating the computation of an object-centric camera pose. In total, 11,530 images were collected as part of 16 minutes of video. A sample of images of a desk are shown in chronological order in Figure 2. For the complete corpus, we plan to collect data from the exploration of 12 rooms, in some cases including two of the same room type (e.g., two different conference rooms, cubicles). We plan to have, on average, about 10 objects of interest with AprilTags in each room.

2.2. Text Annotations

Our aim is to develop four distinct annotation schemes that will be used to construct corpora used for training and testing machine learning models. Each of the annotation tasks are described below. After piloting with colleagues, these tasks will be presented to participants on Amazon's Mechanical Turk. Given our initial data estimate involving twelve rooms with about 10 objects of interest per room, we can estimate about 28 hours of annotation required for a single pass of annotation and, accordingly, about 84 hours for the minimum triple annotation desired for each task.

2.2.1. Task 1: Object Labeling

While the state of the art in object recognition can produce simple captions for images (Xu et al., 2015), first-person context-dependent video summaries require a deeper level of annotation (e.g., features of objects, such as color or spatial orientation in a scene, a notion of which objects and features are salient). To address these needs, our first annotation task is a type of object labeling task in which annotators view an image from the robot's video that has been preprocessed using an existing object recognition tool (e.g., recent convolutional neural network approaches such as Krizhevsky et al., 2012), so that many of the whole objects present in the image are already labeled. The annotator's job will be to correct the existing labels if needed, mark up (by boxing in) and provide labels for any additional whole objects and the identifiable object parts, features and intrinsic properties. What the annotator attends to in the image will be constrained by the instructions, which will guide the annotator to focus on objects that will be notable given a higher-order task (i.e. *information requirement*). For example, an information requirement relevant to disaster relief may be to determine where to establish a communications or headquarters like environment, including determining the location of a working power source. Thus, annotators would have this information requirement in mind when marking up an image, giving them some guidance as to what objects, and which parts and features of those objects, are relevant to this requirement. Thus, for example, annotators would mark the presence of a "lit power button" on the front of a computer.

We will compute inter-annotator agreement to determine annotation reliability (Passonneau and Litman, 1993). Although we will discard any extremely unusual annotations and use agreement to establish the most common label, we would like to keep multiple labels for the same object, part, feature or property in order to capture the range of terms that can be used in descriptions. Dataset 1 will consist of images marked up with boxed-in objects and labels for the objects, as well as the salient parts, features and properties of those objects. This data will augment existing training

Figure 2: Sample sequence of images from initial data collection.

data for the object labeling tool and provide gold-standard data for evaluation.

2.2.2. Task 2: Natural Language Summaries

To determine what kinds of detail to collect about objects, we plan to observe what content people choose to describe when they are tasked with authoring narratives about image sequences. In this task, annotators view a sequence of images drawn from a segment of the robotic exploration. The images will have the markup and labeling from Task 1. The task will be to describe the portion of the path viewed during the sequence as well as the relationships between labeled objects encountered along that journey. Determining the quality of natural language summaries of any type is notoriously difficult (Rankel et al., 2013); thus, we are exploring possibilities including the ROUGE suite of automatic evaluations for summarization (Lin, 2004) and the feasibility of the Pyramid method of evaluation (Nenkova and Passonneau, 2004).

Dataset 2 will consist of the sequence of images and the accompanying narrative descriptions of the path and the relationships between objects encountered during the robot's journey. Recall that the robot builds a 3D map of its environment. We see an opportunity to tie the narrative of the exploration elicited in this annotation task not only to the images from the exploration, but also the robot's path throughout the mapped environment. This would include, for example, associating all of the objects described in a room to the representation of those objects in the map. We are currently exploring the feasibility of automatically establishing the association between the narrative, image sequence, and map.

2.2.3. Task 3: Elicitation of Questions

The third annotation task will support training and evaluation of a system's performance on a relevant-image retrieval task: given a natural language question about an object, return the image most likely to enable a human to determine the answer to the question posed. In this annotation task, annotators will be provided with an information requirement (e.g., determine if there is power in the building). They will also be provided with the narrative description of an environment drawn from Task 2 described above. Given the information requirement and the narrative, the annotator's task will be to provide a question about the environment that is described. The answer to the question provided would allow a person to respond to the information requirement. Thus, for example, one might provide the question, *Is the computer on the desk on the right turned on?* because the answer to this question would help to determine the status of power in the building.

We are currently in the piloting stages of this annotation, trying to determine whether the text description alone is enough for annotators to visualize a space and come up with questions, or if additional information, such as the video stream, is necessary for eliciting such questions. Establishing the quality of crowdsourced annotations is somewhat challenging for this dataset. We are exploring the possibility of including a final step in each annotation task wherein annotators check the appropriateness of another annotator's questions given the same information requirement. Dataset 3 will consist of the information requirement, the elicited questions, and the sequence of images containing objects referred to in those questions. Our findings will inform other annotation task involving complex scenarios.

2.2.4. Task 4: Image Selection

The previous annotation task links a provided question pertaining to an object to a sequence of images depicting that object. This might give a human enough information to answer the question, but it does not provide the training data needed to determine which images are adequate for answering the question, which are inadequate, and which are ideal. For example, given the natural language question *Is the computer on the desk on the right turned on?*, an image of the computer taken at the profile, from behind, or below may not provide the information needed to answer the question. Thus, Task 4 will determine the adequacy of a particular image for answering a particular question.

Annotators will be shown one image at a time, sampled from the set of images associated with a particular question in Dataset 3. The annotator's task is first to indicate whether or not the image displayed enables one to answer the question. Then, the annotator will be asked to select and rank the three most ideal images from the set to answer

the question. Finally, the annotator will be asked to provide the evidence they found in the image to answer the question. For example, they will be asked, *How do you know whether the computer on the desk on the right is turned on?* Participants must answer in the form *Because I can see...*, where they complete this sentence with the feature from the images used as evidence: *Because I can see that the power button is illuminated.* To establish the quality of the annotations indicating which images are adequate and/or ideal for answering a question, we will measure agreement across triple annotations. The final Dataset 4 will consist of the natural language question and the sequence of images from Dataset 3, but with the additional markup of which images are or are not adequate for answering the question and which images are ideal for answering the question, as well as the visual evidence used.

3. Related Work

The original contribution of this dataset is in the annotation of images and video solely from a first person perspective. Other datasets abound for scene and action identification, as summarized for comparison here. Recently, there has been a great deal of research combining previously independent strands of work from language processing and computer vision (e.g., Jiang et al., 2013). Both fields have normally sought to extract relevant features from input and to classify the data accordingly in order to impose some kind of "meaning" on unstructured text and image data. Many of these datasets were built to facilitate object recognition and consist of images of objects and the accompanying natural language label. While initial efforts focused on relatively few object classes (see Everingham et al., 2008), the most recent state-of-the-art dataset, ImageNet (Deng et al., 2009), includes 3.2 million images and covers 5,247 sets of synonymous nouns drawn directly from WordNet (Fellbaum, 1998). Current object recognition training datasets do not, as far as we are aware, provide images with explicit labeling that focuses on object parts, features and properties (e.g., labels on a desk indicating that the top is wood while the legs are metal). Other datasets in this research area provide natural language summaries of a scene captured in an image (e.g., Karpathy et al., 2014) or video (e.g, Rohrbach et al., 2013). While valuable, existing resources do not include sequences of scenes showing movement through an environment from a first-person perspective, and therefore fail to fully meet the needs of training a robot system. In other recent work, visual question answering resources have been developed with the aim of providing natural language answers to questions about particular images (Antol et al., 2015). There is also related work in multimedia question answering (Hong et al., 2012), though we are aware of no prior work that focuses on enabling fine-grained distinction among many images of the same scene based on their ability to "show" the answer to a natural language question about that scene.

4. Unique Contributions, Conclusions

The resulting datasets will provide valuable corpora for the training and evaluation of machine learning systems for object labeling, scene summarization, and image-based question-answering. Furthermore, each dataset fills a void in existing language/vision resources. Specifically, the object labeling is uniquely geared towards the labeling of parts, features and properties of objects rather than whole objects. Furthermore, the scene summaries are unique in that they are first-person narratives of both path and static scene information. Within the realm of image-based question answering, we are establishing which poses are informative for viewing a particular object or types of objects rather than assuming a canonical, human perspective of an object. Additionally, we are eliciting valuable information on the types of evidence present in image that allow humans to answer a particular question. This data is uniquely suited to meeting the challenges of training a robotic system for scenarios such as humanitarian assistance and disaster relief, given that annotators complete their tasks with a higher-order information requirement in mind.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). VQA: Visual question answering. In *Proc. of ICCV*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*. IEEE.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *Int'l Journal of Computer Vision*, 88(2).

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Hong, R., Wang, M., Li, G., Nie, L., Zha, Z.-J., and Chua, T.-S. (2012). Multimedia question answering. *IEEE MultiMedia*, 19(4).

Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., and Shah, M. (2013). High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2).

Karpathy, A., Joulin, A., and Li, F. F. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Proc. of NIPS*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Lemon, O., Bracy, A., Gruenstein, A., and Peters, S. (2001). The WITAS multi-modal dialogue system. In *Proc. of Interspeech*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proc. of the ACL Workshop: Text summarization branches out*.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. *Columbia University Academic Commons*.

Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. In *Proc. of ICRA*.

Passonneau, R. J. and Litman, D. J. (1993). Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of ACL*.

Rankel, P. A., Conroy, J. M., Dang, H. T., and Nenkova, A. (2013). A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. In *Proc. of ACL*.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., and Schiele, B. (2013). Translating Video Content to Natural Language Descriptions. In *Proc. of ICCV*.

Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*.

Building an Audio Description Multilingual Multimodal Corpus: The VIW Project

Anna Matamala, Marta Villegas
Universitat Autònoma de Barcelona
Edifici K, 08193 Bellaterra
E-mail: anna.matamala@uab.cat, marta.villegas@uab.cat

Abstract

This paper presents an audio description multilingual and multimodal corpus developed within the VIW (Visual Into Words) Project. A short fiction film was created in English for the project and was dubbed into Spanish and Catalan. Then, 10 audio descriptions in Catalan, 10 in English and 10 in Spanish were commissioned to professional describers. All these data were annotated at two levels (cinematic and linguistic) and were analysed using ELAN. The corpus is an innovative tool in the field of audiovisual translation research which allows for comparative analyses both intralingually and interlingually. Examples of possible analyses are put forward in the paper.

Keywords: audio description, audiovisual translation, multimodal corpus, multilingual corpus

1. Introduction

Audio description (AD) is an intersemiotic translation in which images are translated into words (Maszerowska et al 2014). These words are delivered aurally to an audience who does not have access to the visuals, mainly the blind and visually impaired but also other users who for various reasons cannot access the visual content. In audiovisual productions, AD is interspersed in the segments where no dialogue and no relevant sounds are heard. Its aim is that the audience can understand and enjoy the audiovisual content only through the audio channel. One could say that AD has been provided informally by sighted people who, for instance, watch television with blind or visually impaired friends or family. Volunteers have also played a key role in making many cultural activities accessible to all. However, AD as a professional access service is more recent, and still non-existent in certain countries (Orero 2007), despite accessibility has been included as a human right in the UN Convention on the Rights of Persons with Disabilities and there is increasing legislation promoting it (see EU's recent proposal for an Accessibility Act).

Guidelines have been developed by standardization bodies, regulators and associations (Matamala and Orero 2013) to help describers in the complex task of translating all the nuances provided by the images into a limited number of sentences or words. A set of strategies designed within the ADLAB project (Remael et al 2016) are a useful tool to help describers in their choices.

Research on AD is also recent and has been integrated within the frame of audiovisual (AV) translation studies (Braun 2008). Investigations on AD have been mainly descriptive, dealing with specific practices such as theatre opera, art, and cinema. Case-studies have approached the analysis of various features in ADs, such as cultural references (Mangiron and Maszerowska 2014), sometimes adopting a contrastive approach (Bourne and Jiménez 2007). More recently, reception research with end users and technological aspects have been tackled. However, AD corpus research has been scarce, and there is still a lot to be learnt concerning both the process of AD and the final product.

This paper presents a corpus of ADs developed within a one-year project (Visuals Into Words, VIW), running from October 2015 until September 2016 under the Spanish Government *Europa Excelencia* funding scheme. VIW's ultimate aim is to create an open access platform that will allow for comparative research on AD, both intralingually an interlingually. To contextualise this project within AV translation studies research, Section 2 summarises the state of the art in corpus research in AD. Section 3 explains the project rationale. Section 4 describes the corpus in its current stage of development, and section 5 defines the corpus annotation procedures. Section 6 puts forward possible corpus exploitations.

2. Previous Work

Most AD research has focused on one film, sometimes expanding the corpus to a few films (Piety 2004). Two relevant exceptions to this trend are TIWO and TRACCE.

TIWO (Television in Words) was a project led by Andrew Salway between 2002-2005 at the University of Surrey (UK) which aimed "to develop a computational understanding of storytelling in multimedia contexts, with a focus on the processes of AD" (Salway 2007: 153). In order to do so, 91 audio description scripts in British English from three major producers of AD were collected, making up a corpus of 618,859 words (Salway 2007: 155). The TIWO corpus allowed Salway to carry out a thorough analysis of the language of AD in English (Salway 2007). It also compelled him to propose some ideas on assisted audio description, and to suggest how AD could be used for keyword-based video indexing.

On the other hand, TRACCE (Jiménez Hurtado et al 2010) was a project led by Catalina Jiménez Hurtado between 2006 and 2009 at the University of Granada (Spain). A corpus of 300 films audio described in Spanish, plus 50 films in German, English and French, were collected. Most of the Spanish scripts came from the film archives of the Spanish blind association ONCE because at the time the corpus was created few commercially available audiodescribed films were available in Spanish. A multimodal annotation system and a specific tool were developed (*Taggetti*) to tag the AD scripts (Jiménez

Hurtado and Seibel 2012: 412). Annotations were created at three different levels: film narrative, camera language, and recurrent grammatical structures in the ADs. The tagging process was carried out manually in one-minute film segments called Meaning Units, which were composed of the AD script and the associated AV content.

Despite the relevance of both projects in AD corpus research, they are not freely available on the Internet, probably due to copyright issues. This is similar to what happens often in other fields of AV translation, where corpora have been created but have faced copyright constraints (Baños, Bruti and Zanotti 2013).

On the other hand, sometimes a significant number of data have been gathered, but they have not been incorporated into a systematic corpus. This is the case, for instance, of the Pear Tree project (Mazur and Kruger 2012) developed within the DTV4ALL project. Participants from different countries collected descriptions of the same film, a clip created for Chafe's (1980) Pear Stories project that contained no dialogue, in order to identify cultural similarities and divergences.

3. Motivation

It is in this context that VIW was born. Inspired by Chafe's (1980) project, and its posterior implementation in AD (Mazur and Kruger 2012), VIW aims to develop a multimodal and multilingual corpus of AD departing from a single stimulus, a short film created *ad hoc* in English, and translated into other languages. This corpus will allow to carry out studies comparing the AD versions produced for one language but also contrasting various languages.

The project is built upon two pillars: on the one hand, it has a strong open access component. All materials will be freely available to the research community, through an open platform that is currently being designed (Creative Commons licence CC-BY-NC-SA). Copyright has been secured through agreements developed specifically for the project, both for the film (in English and in its translated versions) and the ADs created. On the other hand, it aims to be a scalable and expanding project. This means that, although very limited in size in its initial stages, the project is being designed and developed so that it can easily incorporate other languages and inputs provided by external researchers. This will be feasible thanks to a clear documentation of all the processes and the implementation of open access tools and licences.

4. Corpus description

This section describes the corpus considering its current stage of development, but also indicating further developments that will be achieved at its completion. It differentiates between the short film that is at the core of the project and the AD that have been created.

4.1 The Short Film

The short film was commissioned to a film director and produced specifically for the project. To make sure the film would be useful for AD research purposes, a literature review and experts' discussion allowed to identify the key elements that are considered challenges in AD. These included: characters and action, including gestures and facial expressions, spatio-temporal settings, film language, sound effects and silence, text on screen, and intertextual references (Maszerowska et al 2015). The film director was instructed to create a short film with a standard narrative structure, various actions, and at least four characters speaking in English except for one, who would speak another language at least at some point so that subtitles could be added. Further instructions were to include at least three different spatio-temporal settings, and to incorporate some text on screen as well as opening and end credits. The director was told to include in the film narrative at least one sound that could not be easily identifiable, and to show silent passages for artistic purposes. Finally, the film director was made aware that the film would be audio described, hence segments without speech were needed to add the audio description.

It was considered that the film should last a minimum of 10 minutes to allow for research on user engagement, a hot topic in the AD research agenda. At the same time, it was considered that a much longer film would make it more difficult its re-usage in experimental settings and, last but not least, it would also be difficult to support financially. This is why the film director was instructed to create a film between 12 and 15 minutes long. The result is the film "What happens while---", directed by Núria Nia, which lasts 14 minutes, and deals with how different characters envisage time.

Since our aim was to include AD in English, Catalan, and Spanish, a dubbed version of the short film was commissioned to a Barcelona-based dubbing studio. The same translator, dubbing director, and voice talents were used to create both the Catalan and the Spanish version. All three versions will be available from the project website (http://pagines.uab.cat/viw).

4.2 The Audio Descriptions

Ten English AD, ten Spanish AD, and ten Catalan AD were commissioned to professional AD providers. They were requested to generate an AD of the short film following the usual professional standards. They were instructed to send an .mp4 file containing the final audio-video mix plus a time-coded script, without further specifications, over a period of approximately two weeks. Some providers offered the researcher the possibility to make changes to the AD, but it was decided not to intervene in the process and just accept the output as delivered.

As of 12 February, the corpus is made up of the ADs indicated in Table 1[1]. By the end of February, 10 versions per language will be available. An experiment has been planned to gather AD created by AD students to complement the current corpus. This would allow for comparative research between professionals and students.

[1] The full corpus will be available at the end of February and the information will then be updated.

Audio description	#Versions	#Words
English	10	7000
Catalan	7	4200
Spanish	8	4960

Table 1: Number of words and audio descriptions.

5. Corpus annotation

After an analysis of various multimodal corpus analysis tools, ELAN[2] was selected to create complex annotation on the video resources (Sloetjes and Wittenburg 2008). Essentially, ELAN allows linking annotations with their corresponding video files and saves these links in the annotation file. The annotation file is an XML file conforming to the EAF format[3]. ELAN also provides a powerful set of tools to assist video encoding and to perform eventual analysis, hence it was prioritized over other multimodal corpus analysis tools.

Corpus annotation, which is still undergoing, is designed at two main levels:

Linguistic annotations consist of an AD plus a set of dependent layers, where the AD is tied to the time line and the dependent layers are tied to a specific annotation in the audio description itself. Six levels of linguistic dependent annotations have been included, namely: sentences, chunks, tokens, part of speech, lemma, and semantic annotations.

Sentence, chunk and token tiers are simply used to split the AD into smaller parts and, hence, their annotation value is a sub-string of the AD. Lemma, part of speech, and semantic annotation[4] are used to further annotate tokens. Linguistic annotations are automatically encoded using the Standford parser[5] (for English and Spanish) outside the ELAN tool and eventually added into the EAF file. To add Stanford annotations into the EAF files extensive use of the Pympi package[6] was made.

Cinematic annotations are currently being developed and are to be applied to the audiovisual content. They include 'text', 'sound' and 'camera' annotations.

Text annotations encode text on screen, be it the opening credits, subtitles or other text, both added at the postproduction stage or as part of the action (for instance, when a characters reads the contact list on a phone).

Sound annotations are particularly relevant for our research because they have a direct impact on where audio description can be included. They are used to identify silence, music, sound effects, and speech. Since sound annotations may overlap, four different tiers have been defined.

Camera annotations, currently being developed, will focus mainly on scene transitions, which often delimitate different spatio-temporal settings, and also on the cinematic technique of zooming, used to focus the audience attention towards an individual object.

The nature of our primary data together with their corresponding annotation sets give our eventual corpora a rather special character. As illustrated in Figure 1, the corpus contains a single short movie, in three different languages, which has been annotated according to 'filmic criteria' and a set of 30 different 'derived versions', each providing an AD. These 'derived versions' vary in language and provider and are further annotated in linguistic terms. In some way, our corpus constitutes a comparable corpus where up to 30 ADs are aligned against the same annotated timeline.

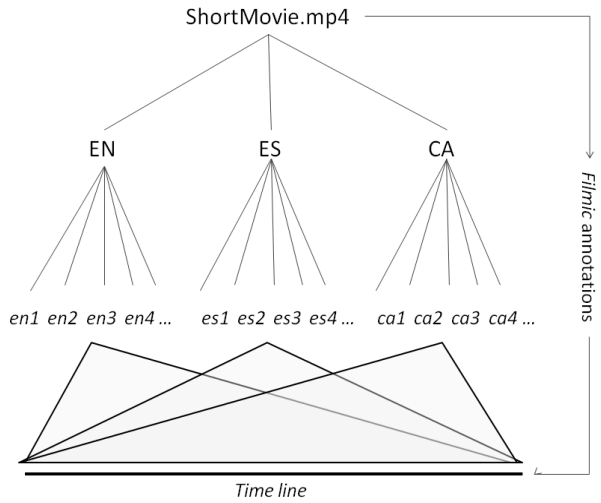

Figure 1: Corpus structure

6. Corpus exploitation

All these annotations allow to perform a wide number of analyses. These may run on a particular file or on a set of files, permitting not only single analysis but also comparative analyses (for example, when comparing among languages or providers). Figure 2 displays two different ADs (one from the UK and another from Canada) in the time line . With this visualization, the researcher can easily see the annotations around a given point of time, quickly identify hot intervals and compare distributions between the two providers, among other features.

[2] See hppt://tla.mpi.nl/tools/tla-tools/elan, developed by the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

[3] The ELAN Annotation Format, also known as the EUDICO Annotation Format Netherlands.

[4] Currently used to classify verbs and adverbs.

[5] Stanford Lexicalized Parser v3.5.2 (http://nlp.stanford.edu/software/lex-parser.shtml).

[6] Pimpy is a package that allows manipulating ELAN and TextGrid files.

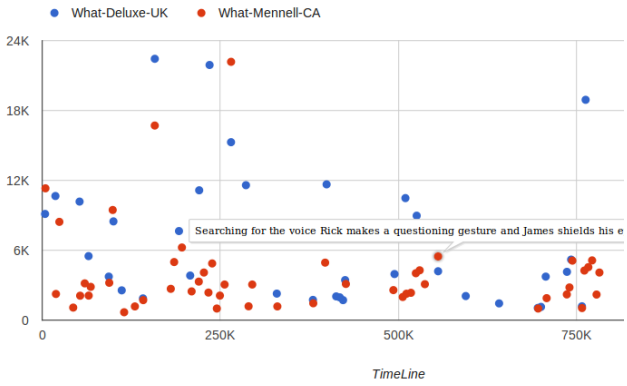

Figure 2: Sample visualization

Similarly, the set of annotations available allows focusing on a single layer or rather mixing different annotation layers. Thus, when considering linguistic layers alone, a variety of calculations can be performed on word frequencies, word distributions (both in time line and among different EAF files), density, etc. Figure 3 shows the number of words/sentences per paragraph.

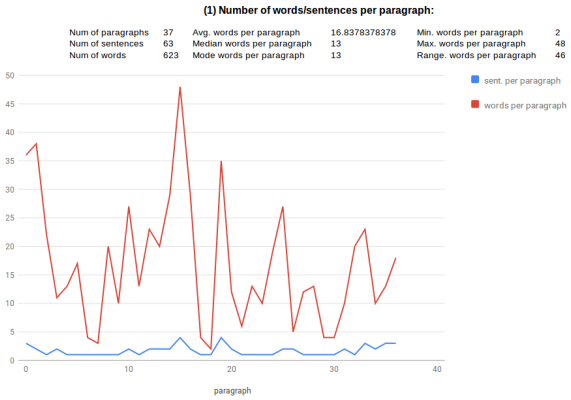

Figure 3: word/sentence visualization

Particularly interesting is the correlation between camera and linguistic annotations. In this case, multilayer concordances allow identifying relevant correlations between significant filmic annotations such as shot transitions or zooms and the ADs. The fact that all annotations are aligned to the same annotated timeline opens a wide range of possibilities.

The web application that is currently being developed provides access to source data and (some) graphical visualizations of that data. Source data include the ELAN annotation files as well as a set of csv files prepared to support the range of experiments and analyses that researchers may define. The graphical visualizations provided aim to explore and exemplify the possibilities of the annotated data available. In this case, Google Charts API is used to generate the charts out of the source data.

7. Conclusions

All in all, this paper has presented an ongoing project whose aim is to develop a corpus of AD created for a single film input. Despite its current limited size, our belief is that it is an innovative resource in terms of audiovisual text types and approach. It is multimodal and multilingual, and allows to compare at various levels how the same visual input is translated into words in different languages and by different describers.

8. Acknowledgements

Project funded by Spanish Ministry within the Europa Excelencia programme (FFI2015-62522-ERC). Anna Matamala is a member of TransMedia Catalonia, funded by Catalan government funds (2014SGR027).

9. References

Baños, R., Bruti, S., Zanotti, S. (2013). Corpus linguistics and AVT: in search of an integrated approach. *Perspectives*, 21(4), pp. 483--490.

Bourne, J., Jiménez, C. (2007). From the visual to the verbal in two languages. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All*. Amsterdam: Rodopi, pp. 175--188.

Braun, S. (2013). Audiodescription research: state of the art and beyond. *Translation Studies in the New Millennium*, 6, pp. 14--30.

Chafe, W. (1980) (Ed.) *The Pear Stories*. Norwood: Ablex.

Jiménez Hurtado, C., Rodríguez, A., Seibel, C. (2010). *Un corpus de cine.* Granada: Tragacanto.

Jiménez Hurtado, C., Seibel, C. (2011). Multisemiotic and multimodal corpus analysis of audio description: TRACCE. In A. Remael, A., P. Orero, & M. Carroll (Eds.), *Audiovisual Translation and Media Accessibility at the Crossroads*. Amsterdam: Rodopi, pp. 409--425.

Mangiron, C., Maszerowska, A. (2014). Strategies for dealing with cultural references in AD. In A. Maszerowska, A. Matamala, & P. Orero (Eds.), *Audio Description*. Amsterdam: Benjamins, pp. 159--178.

Maszerowska, A., Matamala, A., Orero, P. (Eds.) (2015.). *Audio Description*. Amsterdam: Benjamins.

Matamala, A., Orero, P. (2013). Standardising audio description. *IJSEI*, 1, pp. 149--155.

Mazur, I., Kruger, J.-L. (2012). Pear Stories and Audio Description: Language, Perception and Cognition across Cultures. *Perspectives*, 20(1), pp. 1--3.

Orero, P. (2007). Sampling Audio Description in Europe. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All.*. Amsterdam: Rodopi, pp. 111--125.

Piety, P. (2004). The language system of audio description: an investigation as a discursive process. *JVIB*, 98(9), pp. 453-469.

Remael, A., Reviers, N., Vercauteren, G. (Eds.) (2015.). *Pictures painted in words: ADLAB Audio Description guidelines*. Trieste: EUT.

Salway, A. (2007). A corpus-based analysis of AD. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All*. Amsterdam: Rodopi, pp. 151--174.

Sloetjes, H., Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

A Dataset and Evaluation Metric for Coherent Text Recognition from Scene Images

Name of author

Address - Line 1
Address - Line 2
Address - Line 3

Abstract

In this paper, we deal with extraction of textual information from scene images. So far, the task of Scene Text Recognition (STR) has only been focusing on recognition of isolated words and, for simplicity, it omits words which are too short. Such an approach is not suitable for further processing of the extracted text. We define a new task which aims at extracting coherent blocks of text from scene images with regards to their future use in natural language processing tasks, mainly machine translation. For this task, we enriched the annotation of existing STR benchmarks in English and Czech and propose a string-based evaluation measure that highly correlates with human judgment.

Keywords: scene text recognition, machine translation

1. Introduction

Scene Text Recognition (STR) is a subfield of artificial intelligence that has been studied for a long time (Gómez and Karatzas, 2014) with a recent advances achieved by employing deep learning methods(Jaderberg et al., 2014b). With the increasing volume of pictures taken by hand-held devices, scene text (ST) became an interesting potential source of text for processing by Natural Language Processing (NLP) methods. Nevertheless, most of the previously published work strictly focus on recognition of isolated words and do not view the recognized words as utterances that belong into a particular language context. Another drawback of the state-of-the-art STR methods is that the benchmarks usually omit short (mostly function) words, for which they claim there is not enough visual evidence to be recognized. Most NLP methods usually deal with either text that can be split into sentences or directly with text on sentence level. ST on the other hand, consists of rather short chunks, such as proper names, isolated noun phrases or very short sentences. To machine-translate the ST we need to be able to recognize these chunks properly.

The only work mentioning Machine Translation (MT) of ST (Bijalwan and Aggarwal, 2014) we are aware of uses only simple rules for forming coherent text and pass the text to a statistical MT system. No systematic evaluation of the process is given. There exist few mobile applications for MT of ST[1] which very likely work similarly. As far as we know, no one approached this problem more systematically. In the next section, we briefly summarize the state of the art in ST localization and recognition. Section 3. brings a syntactic definition of connected text blocks and introduces a dataset for training automatic coherent text recognition from scene images by enriching existing STR benchmarks. In Section 4., we propose an automatic evaluation metric that allows fast comparison of methods.

[1] *Google Goggles* (`http://www.google.com/mobile/goggles`) and *Bing Translator* for Windows Phones (`http://www.bing.com/translator/phone/`) are the applications we know about.

Figure 1: Examples of ST images from the ICDAR Focused Scene Text Dataset.

2. Scene Text Localization and Recognition

Unlike the well-solved problem of optical character recognition, STR is a more challenging and still not satisfiably solved task. The text is usually placed on heterogeneous background with many distortions including shadows, reflections, and deformations (see the examples in Figure 1). Text extraction from scene images is usually divided into separate steps of text localization and recognition. Even though methods for unbounded recognition (Bissacco et al., 2013; Jaderberg et al., 2014a) exist, the recognition typically uses a limited vocabulary (Roy et al., 2014; Jaderberg et al., 2014b). The state-of-the-art methods are summarized, e.g., in the 2015 ICDAR Robust Reading Competition results (Karatzas et al., 2015).

3. Coherent Text Reading

Our goal is to indetify blocks of coherent text which can be further used in NLP tasks. We thus want to find *the minimum coherent text blocks closed on syntactic dependencies* (as perceived by an annotator, not automatically computed). Our original idea of the coherent text blocks was semantically motivated. We observed that ST frequently has a hierarchical nature. Signboards often contain lists of offered goods or services (coordinated on the same level), with a name of a venue as a kind of headline of the list on which the items depend. This hierarchy induces a natural order in which readers read the words. Annotating this would be

Figure 2: Example of an image with focused ST (left) and image with incidental ST (right). The word bounding boxes are highlighted by colorful boxes.

very laborious. Moreover, most of the hierarchies in the existing STR data are very flat, so such complex annotation would not pay off.

This is related to a problematic syntactic phenomena for the block definition which are coordinations indicated entirely by visual means where a coordination token is missing. Other ellipsis could be identified also on the pragmatic level (e.g., missing 'this shop offers:' on a signboard).

To avoid these problems we disregard all dependencies that are not explicitly present in the text. Unlike the standard STR benchmarks, we do not rely on the visual evidence only and also include cases where the text is obvious from the language context.

3.1. Original STR Data

The most frequently used benchmarks in STR come from the ICDAR Robust Reading Competition (Karatzas et al., 2015). For every competition, the annotation and evaluation protocol slightly differ. In the 2015 competition, all words in the images were localized in quadrilaterals and most of them were accompanied with a transcription. Words that are not readable or are shorter than 3 characters are marked as "not-care" words. *Focused ST* and *incidental ST* (see Figure 2 for examples) are distinguished as separate categories.

In the focused ST dataset, the main purpose of taking the pictures was the text. The pictures usually capture signboards and notices from an urban environment together with a few book covers and signs of electronics. The dataset consists of 229 training and 223 test images. On average, there are 6 words in an image out of which less than 3 are the "not-care" words. Most of the text is in English, with a few images containing signboards with a text in German.

The dataset of incidental ST consists of 1,000 training images and 500 test images taken in streets, shopping centers, and public transport of Singapore. The images capture complete urban scenes with a lot of text which often suffer from being out of focus and motion-blurred. There is, on

	pilot		final	
dataset	F	acc	F	acc
English focused	.820	.705	.943	.917
English incidental	.533	.190	—	
Czech focused	.853	.600	.962	.900

Table 1: Average inter-annotator agreement for both the pilot and final annotation.

average, 12 words in each image out of which 7 are "not-care" words. Most of the text is in English with some signs in non-Latin scripts which are localized but not transcribed.

The benchmarks only expect words from certain vocabularies to be recognized. For that purpose, sets of 50, 1k and 90k words are provided. Even though, the biggest lexicon may seem big enough for English, it may not be sufficient for languages with rich inflection or compounding. In addition, we use 81 images of Czech focused text (Hadáček, 2014) with 16 words per image.

Apart from the mentioned datasets, there exist other datasets worth mentioning. The *KAIST Scene Text Database* (Jung et al., 2011) consists of 3k images with focused texts in English and Korean. The *NEOCR* dataset (Nagy et al., 2012) is a set of 659 real world images with more than 13k words annotated on line level instead of word level.

3.2. Annotation Process

We annotated the coherence by explicitly marking chains of words in the images. Initially, we did a pilot annotation of 20 images from both ICDAR 2015 focused and incidental datasets and the Czech focused text dataset. Five annotators were provided with a simple definition of the task with little further details. They were asked to add transcription of "not-care" words if possible and to mark cases where a single word has been falsely split into multiple bounding boxes. An example of the annotation is in Figure 3.

We measured the inter-annotator agreement by *mutual accuracy* defined as a proportion of images that have been equally annotated and *mutual F-score* defined as a harmonic mean of the precision of the first annotator given the second one and vice versa. Values are tabulated in Table 3.

During the pilot annotation we experienced some problems with guessing the text. Different annotators set themselves different thresholds when they are certain about a word. The incidental text dataset was acquired in Singapore with a high density of shops. One annotator familiar with the fashion brands was able to transcribe much more signs than the others. Another annotator admitted he searched the Internet to find unreadable titles of books whose covers were in the dataset claiming that the image provided him with enough information to find out what the rest of the text is. The low agreement in the incidental dataset was mostly because the annotators were inconsistent in deciding what is readable in the images and what is not. With 10 seconds per word on average, the incidental text took more than twice as long as in the case of focused text annotation.

Based on the pilot annotation, we decided to only annotate the focused ST images. The annotation guidelines were refined to cover the most frequent inconsistent cases. These

image

STR annotation

```
35,120,390,120,390,222,35,222,HSBC
39,256,131,256,131,295,39,295,The
151,259,327,259,327,298,151,298,world's
350,261,461,261,461,300,350,300,local
484,262,602,262,602,301,484,301,bank
```

coherent text

```
1(The),2(world's),
3(local),4(bank);
0(HSBC)
```

Figure 3: An example of annotation of an image with focused ST. Numbers preceding the brackets refer to the line numbers in the original annotation, blocks are separated by semicolons.

were: a headline is a separate chunk; if a new line in the text is a substitute for a punctuation mark, it is a block separator; an address should be segmented as on an envelope; ignore characters which are not text (e.g., *P* for parking place); searching for additional knowledge is not allowed. An ex-post standardization was done on the annotation of rare punctuation (trade-marks, bullets, and vertical bars) increasing the mutual accuracy by 10 percentage points. The inter-annotator agreement on the final annotation is tabulated in Table 3.

In total, 81 images of Czech and 452 images of English focused ST were annotated. The images contain 3.6 blocks per image on average with the average length of 3.3 words. The images with the Czech focused text contain on average 4.9 blocks per image with the average length of 2.7 words. The dataset is relatively small. We expect the training part of the ICDAR Focused Scene Text can to be used for training postprocessing of the STR results. The Czech data and the test part of the ICDAR dataset will be used for testing.

4. Evaluation Metric

For training and comparing automatic methods for coherent text recognition, an automatic evaluation measure is needed. The standard STR evaluation metric (Karatzas et al., 2015) is a conjunction of the localization and string correctness. With coherent text recognition, we would like to have a measure that captures how comprehensible text would be if we did not have an access to the image. We believe we can disregard the text location and evaluate the transcription purely based on text similarity because for further text processing the text location does not matter at all. We first tried to explore the human perception of the recognition errors and based on that we designed an evaluation measure. We then explored different configurations of the measure and selected one that agreed the most with the human judgment.

4.1. Experiments

We asked annotators to evaluate erroneous transcriptions of the ST. It was done by three annotators who participated in the pilot annotation (were familiar with the task) but not in the main annotation (were not biased by already having seen images).

We generated two artificial erroneous transcriptions for each of the images that were previously unseen by the annotators. They were asked to imagine they are receiving the blocks in a random order and should translate them to a different language without seeing the image. Then they chose the one they think would lead to better translation.

error type	weight	
	human	machine
character insertion	12.8	3.6
character deletion	12.4	5.9
character substitution	12.8	6.0
block join	20.5	34.8
block split	24.0	34.7
block permutation	17.6	15.0

Table 2: Comparison of the estimated error weights for human annotators and the best fitting automatic measure.

The transcription errors were: character insertions, deletions, and substitutions, joining two blocks, splitting a block into two, permuting words within a block. The edit operations were sampled randomly from the distribution of edit changes obtained from running the TextSpotter STR tool (Neumann and Matas, 2012) on the same dataset. The annotators evaluated three different pairs of transcriptions for each image. One third of them was common for all annotators and was used to measure the inter-annotator agreement. The average-pairwise agreement was 0.670 with Cohen's kappa equal to 0.341.

To roughly estimate the importance of different error types for the annotators, we can view their decisions as a result of a linear combination of the error counts in each image. We do the estimation by fitting a logistic regression model. Normalized weights obtained from the model are tabulated in Table 2.

The model shows that the annotators consider joining or splitting blocks to be more serious errors than the character edit operations that all received similar weights.

4.2. Automatic Measure Description

Because the blocks can be recognized in a random order, we need to match the transcription and reference chunks before measuring their similarity. Expecting a reasonable quality of the underlying STR, we can match the reference with the recognition using entirely by the string similarity, disregarding their spacial position.

Formally, let $\mathbf{b} = (b_1, \ldots, b_n)$ be a machine-generated blocks, $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_m)$ its reference blocks and G a complete bipartite graph with $\mathbf{b}$ and $\hat{\mathbf{b}}$ its partite sets. The edges are weighted by string similarity $\mathrm{sim}(b_i, \hat{b}_j) \in [0, 1]$. A chunk matching $M \subset \mathbf{b} \times \hat{\mathbf{b}}$ is obtained as the minimum weighted maximum bipartite matching (Munkres, 1957) in the one-to-one case or as a minimum weighted edge cover (Schrijver, 2003) in case of many-to-many matching. We

	MM	MWEC
norm. Levensthein	.723	.722
Marzal-Vidal	**.726**	.724
Jaro-Winkler	.701	.715
PER	.645	.646
3-gram prec.	.685	.682
4-gram prec.	.696	.690
5-gram prec.	.696	.688

Table 3: Average agreement of the different configurations of thea automatic measure with the human judgment.

define the evaluation measure as:

$$m(\mathbf{b}, \hat{\mathbf{b}}) = \frac{\sum_{(b_1,b_2)\in M} 1 - \mathrm{sim}(b_1, b_2)}{\max(|\mathbf{b}|, |\hat{\mathbf{b}}|)} \qquad (1)$$

We explored the following string similarity measures: normalized Levenshtein distance, Marzal-Vidal distance (Marzal and Vidal, 1993), Jaro-Winkler distance (Winkler, 1990), position independent word error rate (Tillmann et al., 1997), and character n-gram precision as defined by Papineni et al. (2002). The last two measures are non-symmetric. Marzal-Vidal distance is the only one satisfying the triangular inequality, thus combined with the maximum matching algorithm, yields a distance metric.

4.3. Agreement with Human Judgment

For each pair of transcriptions presented to the annotators, we compute the similarity with the ground truth transcription. We measure the agreement as a proportion of cases when the annotator voted for the transcription with higher similarity score with the annotation. Surprisingly, the agreement with the automatic measures is higher that between the annotators themselves. It may be because the annotators must have picked randomly in cases when it was hardly distinguishable which transcription is better. The values are tabulated in Table 3.

The asymmetric measures lead to approximately the same agreement as the annotators reached with each other. A higher agreement was achieved by using the similarity measure that counts the edit operations which corresponds to the finding that the annotators attributed approximately the same weight to all character-level edit operations. The best underlying similarity measure is the Marzal-Vidal distance.

The best measure also appears to weight the importance of the error types more similarly to the human annotators (see Table 2), although it underestimates character edit operations.

5. Conclusions & Future Work

We introduced a task of coherent text recognition from scene images, enriched the existing STR benchmarks for this task, and proposed an automatic evaluation metric. Although the measure disregards the localization and is based entirely on text similarity, it achieves high agreement with human judgment. As a future work, we would like to machine-learn automatic procedures for this task.

6. References

Bijalwan, D. C. and Aggarwal, A. (2014). Automatic text recognition in natural scene and its translation into user defined language. In *PDGC 2014*, pages 324–329. IEEE.

Bissacco, A., Cummins, M., Netzer, Y., and Neven, H. (2013). PhotoOCR: Reading text in uncontrolled conditions. In *ICCV 2013*, pages 785–792. IEEE.

Gómez, L. and Karatzas, D. (2014). Scene text recognition: No country for old men? In *Computer Vision-ACCV 2014 Workshops*, pages 157–168. Springer.

Hadáček, J. (2014). Detection and recognition of diacritical and punctuation marks in real-world images. Master's thesis, Czech Technical University, Prague.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014a). Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014b). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

Jung, J., Lee, S., Cho, M. S., and Kim, J. H. (2011). Touch TT: Scene text extractor using touchscreen interface. *ETRI Journal*, 33(1):78–88.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V., Lu, S., Shafait, F., Uchida, S., and Valveny, E. (2015). ICDAR 2015 competition on robust reading. In *ICDAR 2015*, pages 1156–1160, Aug.

Marzal, A. and Vidal, E. (1993). Computation of normalized edit distance and applications. *IEEE PAMI*, 15(9):926–932.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Nagy, R., Dicker, A., and Meyer-Wegener, K. (2012). NEOCR: A configurable dataset for natural image text recognition. In *Camera-Based Document Analysis and Recognition*, pages 150–163. Springer.

Neumann, L. and Matas, J. (2012). Real-time scene text localization and recognition. In *CVPR 2012*, pages 3538–3545, California, US. IEEE.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318. ACL.

Roy, U., Mishra, A., Alahari, K., and Jawahar, C. V. (2014). Scene text recognition and retrieval for large lexicons. In *ACCV 2014*.

Schrijver, A. (2003). *Combinatorial Optimization - Polyhedra and Efficiency*. Springer.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, page 2667–2670, Rhodes, Greece.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.

Thinspiration and Anorexic Tweets

Ian Wood
Insight Centre for Data Analytics
National University of Ireland, Galway
ian.wood@insight-centre.org

Abstract

The online presence of anorexics and other people with eating disorders, the "pro-ana" (pro-anorexia) movement, has received much attention both in the media and in eating disorder research (McColl, 2013), with much contention about harmful and beneficial effects on the individuals who take part in the community. Several micro-blogging platforms are used by this community, notably including Tumblr, Instagram and Twitter. I present a collection of Tweets containing a selection of hash tags used by people with eating disorders that were collected over a 3 year period from November 2012. Images are a very prominent and important part of the collected data, with 71% of the tweets containing images and many of those tweets containing few or no words. In this demonstration, I will present some overall statistics and initial analyses of this data set as well as opportunities for enriching the data through detecting image features specific to this context, providing examples of common image types and features and an explanation of their relevance as symbols of the "thin ideal" and as indicators of the psychology of tweet authors and how that relates to eating disorder research. A multimodal analysis of this data, with image features combined with text, promises to yield greater and more informative insights into the Twitter eating disorder and thinspiration community than text analysis alone.

Keywords: multimodal corpus, longitudinal, anorexia, pro-ana, Twitter, images

1. Data Collection

Data was collected from Twitter over a 3 year period from December 2012. A search query consisting of hash tags deemed to be used almost exclusively by people with eating disorders was constructed through an iterative process of collecting tweets for several days on the current set of tags, then manually identifying tags in the collected tweets that appeared to be used by people with eating disorders. Tags that had wider usage were excluded with the exception of *#thinspriation*[1] and related tags — these tags are used also by people concerned about their appearance, but not exhibiting behaviours indicative of eating disorders. Their extensive usage within the eating disorder community and relevance made them important to retain. After 3 iterations of this process, a stable set of tags was identified. Tweets were collected on the final set of tags for a longer period and no extra tags were identified.

The resulting set of tags were:, *#proana*, *#promia*, *#anasisters*, *#bulemia*, *#bulimic*, *#ednos*, *#edproblems*, *#hipbones*, *#thingsanataughtme*, *#thinspiration*, *#thinspo*, *#abcdiet* and *#thighgap*. Further details of the data collection methodology can be found in (Wood, 2015a).

2. Overall Data Statistics

The collected data contains approximately 1.4 million tweets, user profile snapshots from over 300,000 users and over 200,000 unique image ids (though many are in fact duplicates). Hash tags related to *#thinspriation* dominate the data, with 73% of tweets. Retweets and images also account for a significant portion, with 57% of collected tweets retweets and 71% containing images. Thinspiration tweets account for 89% of the images, 80% of the retweets contain images and 76% of retweets contain thinspiration related tags.

3. Candidate Image Features

As noted above, 89% of collected images were tagged with variations of the *#thinspriation* tag. It is not surprising then that collected images are predominantly young and/or beautiful women, with frequent appearances of celebrities. Selfies (photos of the person tweeting, taken by themselves) are common, though many exclude the face and head, and images of body features used as metrics of thin-ness are common (features such as "thigh gaps", flat or spoon-shaped stomachs, arms thin enough to close the fingers around etc. . .). There is a significant minority of somewhat disturbing images indicating body features such as protruding ribs, skin scratches and emaciated women. Other significant minorities, though not clearly linked to eating disorders, include sexy images (often tagged as such) and "*#fitspo*" or "*#fitspriation*" — images intended to inspire exercise and fitness as a path to beauty and desirability. Many images contain text, often overlaied over a photograph or drawing, and photo-collages are also not uncommon.

Preliminary image analysis work was able to identify the presence of faces in images with over 80% accuracy. The presence or absence of faces was considered important as posting an image of part of a body without the face indicates detachment between the the posters self and their body — the body seen as external to their sense of self, which is a known characteristic of many people with eating disorders.

Other features indicated as worthy of investigation include: celebrity images (via search on image databases such as TinEye[2]), body sections symbolic of the thin ideal (e.g.: thigh gaps, stomachs, protruding ribs, collar bones or hips), the presence of text on the image, indoor and outdoor scenes, emotion detection from faces and/or body stance, image collages and split images (typically before/after images). Many non-celebrity images in the data have been

[1] Images intended to inspire the quest for the perfect, thin body.

[2] http://tineye.com/

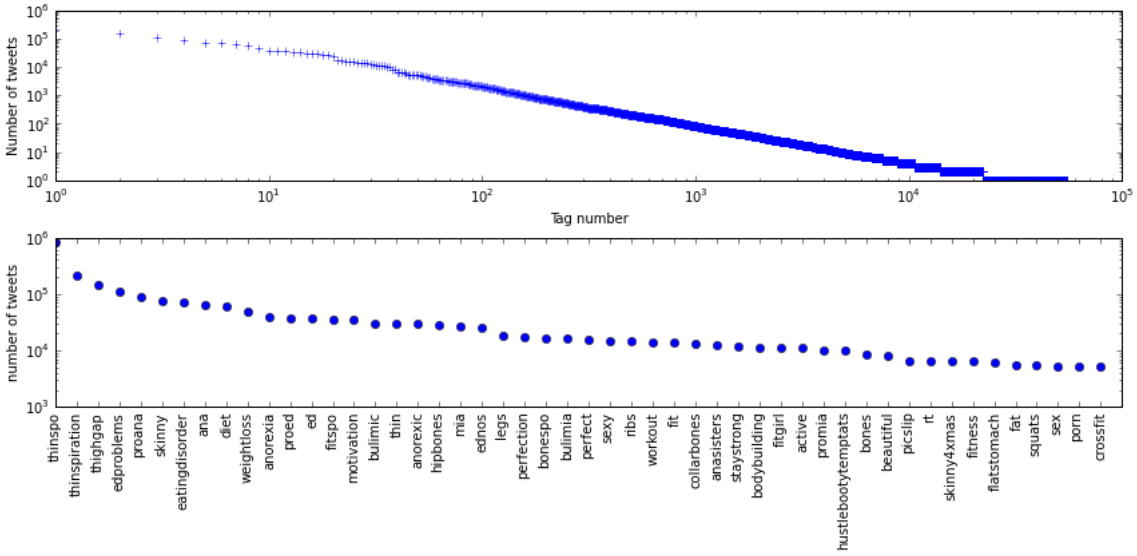

Figure 1: Tag distribution and most frequent tags.

extensively reproduced on the internet. It would be interesting to compare image duplication within the pro-ana twitter data to internet-wide indications obtained from services such as TinEye.

Though reproduction of example images here is not possible (due to privacy and copy write concerns), a search on Twitter with the hash tags above will quickly provide examples.

4. Dynamic and Network Aspects

Alongside tweets containing the selected tags, snapshots of the tweeting users profile and friend and follower lists were collected shortly after each tweet — approximately 1.6 million profile and list snapshots each were collected. In this way, a record of changes to the user profiles and follower networks of frequent tweeters was obtained. A statistical survival analysis is underway to determine factors such as language and hash tag usage and user profile characteristics that correlate with link formation and dissolution. Usage of image types would be another interesting factor to consider in this analysis.

5. Community Identification

A study of communities within the follower network and their usage of textual topics obtained Latent Dirichlet Allocation found strong connections between detected communities and particular topics (Wood, 2015b). This result has twofold implications: first, that the community detection methodology employed appears to have found actual social communities acting as forums for their own topic-specific discussions, and that something of the nature of those communities can be inferred from the associated topics.

Such detected communities can be used both as a filter to narrow further investigations (excluding parts of the data not apparently related to eating disorders, for example) and as a way to associate image features with particular communities and thus gain more directed insights into the behaviour and characteristics of those communities as well as their role in the lives of community members.

6. Conclusion

This demonstration presents an image rich data set of interest to the study of eating disorders and their manifestations as online communities, highlighting the opportunities for multi-modal analysis (image, text and network) in order to seek insights into those disorders and the people that carry them. Such insights have significant public health impact, and may help to address a significant social problem faced by many countries in the world today.

7. Acknowledgements

This work was funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/ 12/RC/2289 (INSIGHT) and the European Union supported project MixedEmotions (H2020-644632). Data collection and analysis were performed with the aid of the Australian National eResearch Collaboration Tools and Resources (NeCTAR) Research Cloud.

8. Bibliographical References

McColl, G. (2013). Anorexia underworld. *The Age*.

Wood, I. (2015a). A case study of collecting dynamic social data: The pro-ana twitter community. *Australian Journal of Intelligent Information Processing Systems*, 14(3).

Wood, I. D. (2015b). Community topic usage in social networks. In *CIKM Workshop on Topic Modelling, Post Processing and Applications*, TM '15, pages 3–9, Melbourne, Australia. ACM.

Collecting a comparative corpus of human-machine, wizard of oz, and human-human chat dialogues

Emer Gilmartin, Ketong Su, Yuyun Huang, Christy Elias, Benjamim R. Cowan, Nick Campbell
Speech Communication Lab
Trinity College Dublin
gilmare@tcd.ie, nick@tcd.ie

Abstract

We describe the design and collection of a corpus of human-human, human-machine and human-wizard of oz interactions. The interactions consist of short conversations with two components - friendly, sometimes teasing chat and a guessing game. The corpus was collected in the form of controlled experiments, with each subject taking part in an interaction with an automatic dialogue system, with a wizard of oz version of the same system, and with another human. The corpus is being used to investigate whether human provided timing is preferable in this conversation genre to automatic 'trailing silence' and the knowledge gained will be used to inform the ongoing development of the CARA and JOKER dialogue systems. A parallel corpus in French was collected by our partners and the two corpora will be used for comparative studies.

Keywords: multimodal corpora, casual conversation, human-machine interaction

1. Introduction

Humans talk for many reasons – much of the activity of our every day lives is mediated by speech. Applications where humans talk to machines have until recently focussed on transactional or task-based domains, such as travel bookings, sales and order fulfilment, or direction giving. In recent years there has been a growing interest in creating applications which converse with humans in a friendly or social manner. Among the challenges which these applications pose is that of dialogue timing - deciding how long an interspeaker gap to leave before the system talks after the interlocutor has finished, or indeed whether to overlap or start speaking before the interlocutor has stopped. Although overlaps are common in human dialogue, our current work focusses on interspeaker gaps.

In any dialogue system, the timing of a response is affected by many factors. First, the system must decide that the interlocutor has in fact stopped talking; this 'endpointing' is generally achieved by setting a trailing silence threshold. If a silence occurs which exceeds this threshold, the system concludes that the interlocutor has finished. There are other delays involved due to factors such as processing time for speech recognition, response selection and synthesis. All of these result in a gap before the system begins speaking. This gap can be lengthened by introducing a delay before speech is triggered. For an interaction to feel natural, it is likely that gap lengths approaching those in human-human talk in a similar context to the desired human-machine interaction would be preferable. In human-human speech research, there have been several studies on the parameters of these gaps, for a comprehensive review is given in Heldner and Edlund (2010). They report that different studies have found differing distributions of gap length, and that gap length can vary with stress levels and cognitive load involved in interaction, modality - whether interlocutors have eye contact, and the type of interaction. They also mention the possibility that speakers adapt to each other's gap length.

Many corpora of human spoken interaction comprise phone conversations (e.g. Switchboard (Godfrey et al., 1992)), artificial tasks (e.g. Maptask (Anderson et al., 1991)), and real or staged workplace meetings (e.g. ICSI (Janin et al., 2003), AMI (McCowan et al., 2005)). While these data have proven very valuable for research into spoken task-based interaction, they may not provide accurate models for friendly social interaction. In addition, the thrust in early dialogue system design was toward task-based systems. In these systems, where the lexical information transferred drives the progression of the dialogue and success depends on task completion, prosodic factors such as gap length may not be a major factor in user satisfaction. However, timing and the 'feel' of a conversation may well take on a greater role in the success of a casual or social conversation. Casual social conversation is described as 'talking just for the sake of talking'(Eggins and Slade, 2004), and subgenres include smalltalk, gossip, and conversational narrative. Aimless social talk or 'phatic communion' has been described as an emergent activity of congregating people, and viewed as the most basic use of speech (Malinowski, 1923). Researchers in fields including anthropology, evolutionary psychology, and communication have theorized that such talk functions to build social bonds and avoid unfriendly silence, rather than exchange linguistic information or express thought as postulated in much linguistic theory. Instances of these views are found in the phatic component in Jakobson's model of communication (Jakobson, 1960), distinctions between interactional and instrumental language (Brown and Yule, 1983), and theories that language evolved to maintain social cohesion through verbal grooming (Dunbar, 1998). As the goal of this conversation is to 'pass the time' rather than to exchange information in order to complete a clearly defined short term task, we speculate that the success of such interactions may depend more on factors such as interpeaker gap length. For example, a short gap may make the conversation seem rushed or uncomfortable, whereas an overlong silence might give the impression of

boredom. To investigate these questions, we are currently experimenting with dialogue systems which chat to, joke with, and indeed tease interlocutors, and are interested in whether and how gap length influences users' impressions of these systems. As part of the JOKER project, we have been improving on our CARA dialogue system which engages in a chat with interlocutors.

2. CARA Dialogue System

The initial Python-based version of the system was adapted from our earlier HERME social talk system (Han et al., 2012). In view of the multimodal nature of the project, it was decided to build a more elaborate Java-based system which could be expanded to incorporate more sophisticated functionality. As an initial step, work was carried out attempting to update and adapt the SEMAINE system, (Schroder, 2010), for use as a social talk system. It was found that this task of re-engineering would be extremely time-consuming and might not result in the type of system we needed to address our research goals of creating realistically timed social talk. It was thus decided to build a custom system Java system. The latest iteration of the system uses CMU's Sphinx ASR and Cereproc's Caitlin Irish accented voice, but is configurable to use other ASR and TTS applications as desired. A WOZ system has been integrated into the Java system which allows a WOZ user interface to be generated automatically for any dialogue flow loaded into the system. This allows us to run experiments contrasting user experience when timing is provided automatically versus when a human times response initiation. Both the automatic and WOZ systems are browser-based and can be run remotely, removing the possible distractions of a researcher in the room where the interaction is taking place. Using this system, we have collected a corpus of human-machine, human-WOZ and human-human dialogues, as described below, which will allow us to perform within-subject analyses of timing in the different modalities.

3. Corpus design and data collection

The work described here forms part of the JOKER project, which aims to build dialogue systems with social communication skills including humor, empathy, compassion, charm, and other informal socially-oriented behavior. For the project, team members in France and Ireland built French and English speaking dialogue system prototypes. The domain chosen for the dialogue was dyadic social talk about food. A short interaction with two phases was devised which was implemented in both languages. The first phase was a 'blague' or 'joshing' stage where the system engaged the user in a short chat about themselves and about food, while producing puns and teasing. The second phase was a guessing game, where the user attempted to guess the system's favourite dish. The first collection of automatic and WOZ recordings for the French speaking system are reported in Devillers et al. (2015), while the data collection for the English speaking system is described below. The data recordings were designed as controlled experiments. There were two conditions, human-machine and human-human.

In the human-machine condition, the same dialogue by each participant was performed in two separate sessions – one with the system running in automatic mode, and another in WOZ mode where a human chose WHEN to make the next utterance but not WHAT to say. Both the automatic and WOZ conditions contained social teasing and guessing game stages. The WOZ and automatic conditions were balanced to prevent any confounding order effects, and held a week apart in order to reduce any priming effects produced by performing similar dialogues in sequence. The human-human condition was added to allow us to extend our within-subject experiments to contrast human and human-machine social talk. The content of the human-human sessions was designed to be as similar as possible to the human-machine sessions, with pairs of naive subjects instructed to chat together and then to play 'Guess my favourite food'. The collection has resulted in a database of recordings of each subject in all three conditions. In addition to participating in each of the three dialogue conditions, in both the French and Irish collections, participants filled out questionnaires on sense of humour and assessments of dialogue quality after each interaction with the system (automatic or WOZ).

In the first round of recordings, there were 16 participants. Participants were recruited by advertising in two Dublin universities, Trinity College and University College Dublin, for native English speakers. The 16 participants comprised 7 male and 9 female native speakers of English ranging in age from 18 to 40, all living in Ireland. None of the participants had any connection with speech and language technology. The recordings were held over a two-week period in a quiet room at the Speech Communication Lab, Trinity College Dublin. Each participant came for two sessions, thus completing all three conditions.

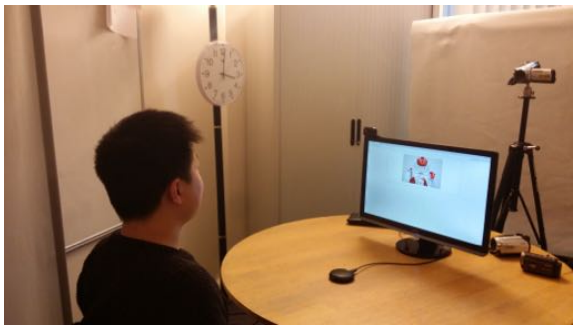

Figure 1: Human Machine Setup

For the human-machine conditions, the subject was seated at a table opposite a screen showing an image of a robot as in Fig. 1. It should be noted that this face to face configuration may not be totally natural but was necessary in order to collect video of the subject's face suitable for later analysis. The subject was fitted with a radio microphone on their chest for near field recording. There were two video recordings made - one using a HD video camera fixed on the subject which also collected audio, and a webcam collecting a

'birds eye' view from above. Audio was also collected by a USB microphone on the table between subject and screen. The subject was asked to wait until the system spoke and then to respond naturally. The interaction was controlled from an adjacent room. In the WOZ case, the same experimenter controlled the timing for all participants. In the WOZ condition, the experimenter could not control WHAT was said by the system, but only press a button to play the next utterance.

For the human-human recordings, pairs of subjects sat opposite each other, with cameras facing each of them as in Fig. 2

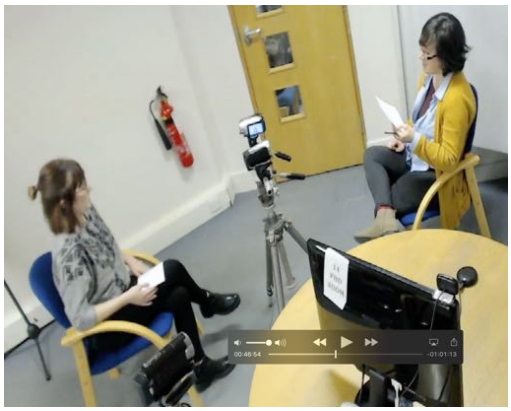

Figure 2: Human Human Setup

Both subjects wore radio microphones affixed to their chests. They were introduced to each other and told they would be left to chat for a few minutes but no instructions were given on what to talk about. Care was taken to ensure that all participants understood that they were free to talk or not as the mood took them. After 10 minutes, the experimenter knocked on the door of the room, entered, gave the participants cards with instructions for the guessing game, took questions if necessary, and left the room. When the game was won by one of the participants, they knocked on the door to signal to the experimenter.

The recordings from the first data collection are being processed and annotated, and will be made available as a language resource in the future. A second cycle of recordings is planned, which will bring the total of subjects recorded up to 32.

4. Conclusion

We have described the collection of a corpus of human-machine and human-human data, which mirrors an earlier collection of french language data. The data will initially be used to analyse whether users prefer human timing of dialogue to automatic, to investigate the parameters of inter-speaker gap length in similar conversations in human-human and human-WOZ conditions, and to use this knowledge to design and implement improved timing modules for dialogue systems performing casual social talk. We hope the data will provide insight into the timing needs of a social system and will help in the transition to an improved version of our chat system. It is also hoped that the recordings will be of use to other researchers.

5. Acknowledgements

This work is supported by the JOKER project and by CNGL.

6. Bibliographical References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Brown, G. and Yule, G. (1983). *Teaching the spoken language*, volume 2. Cambridge University Press.

Devillers, L., Rosset, S., Duplessis, G. D., Sehili, M. A., Bechade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., and others. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 348–354. IEEE.

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Harvard Univ Press.

Eggins, S. and Slade, D. (2004). *Analysing casual conversation*. Equinox Publishing Ltd.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520.

Han, J. G., Gilmartin, E., DeLooze, C., Vaughan, B., and Campbell, N. (2012). The Herme Database of Spontaneous Multimodal Human-Robot Dialogues. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, October.

Jakobson, R. (1960). Closing statement: Linguistics and poetics. *Style in language*, 350:377.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., and Stolcke, A. (2003). The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–364.

Malinowski, B. (1923). The problem of meaning in primitive languages. *Supplementary in the Meaning of Meaning*, pages 1–84.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., and Karaiskos, V. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

Schroder, M. (2010). The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Advances in human-computer interaction*, 2010:2.

Search and Annotation Tools for Heritage Language Spoken Corpora

Kristin Hagen, Janne Bondi Johannessen, Anders Nøklestad, Joel Priestley
The Text Laboratory, University of Oslo / P.O. Box 1102
Blindern, 0317 Oslo, Norway
E-mail: kristiha@iln.uio.no, jannebj@iln.uio.no, anders.noklestad@iln.uio.no, joeljp@iln.uio.no

Abstract

Spoken corpora come in many shapes and sizes, often with their own, unique requirements. This demonstration looks at work with heritage languages, using the Corpus of American Norwegian Speech to show some of the challenges and design considerations. Following a brief account of data collection, focus will be on automatic annotation and providing multimodal access to the corpus.

Keywords: heritage language, multimodal, spoken corpus, transliteration

1. Introduction

The Corpus of American Norwegian Speech (CANS) is a spoken corpus of heritage language, built from informal conversations with and between 50 informants in the USA and Canada. The informants were all heritage speakers, being second, third or fourth generation immigrants, and having learned Norwegian at home. The average age of the informants is high, with the majority being between 70 and 90, reflecting the fact that heritage Norwegian is a dying language. The conversations were recorded between 2010 and 2013, on location, in 22 towns spread throughout the Midwest, the Pacific Northwest and the Prairie Provinces. They comprise some 180,000 phonetically transcribed words. Automatic annotation techniques were subsequently applied to each transcription, to provide corresponding orthographic and morphological layers. The demo will illustrate the resulting, searchable corpus with its search tool Glossa.

2. Data Collection

The conversations are natural, spontaneous exchanges, both between pairs of informants known to one another, and between informant and researcher. They were captured on video, both facilitating transcription and enriching the corpus content. The use of lavalier/lapel microphones provided reliable, unobtrusive audio, regardless of surroundings.

A range of metadata pertaining to the informants was gathered and stored for use in the corpus. Along with age, gender, date of birth, etc., other useful variables were recorded, such as language of instruction at school, contact with Norway, birth place of Norwegian ancestors and whether Norwegian was the informants' mother tongue.

3. Annotation

The first batch of recordings was transcribed using Transcriber. However, owing to requirements regarding multi-tiered transcription, a switch was made to Elan. Elan allows distinct types of annotation to be placed on separate, dependent tiers.

3.1 Transcription

The initial level of annotation is a coarse-grained phonetic transcription. It is the only part of the corpus that is produced manually. The IPA standard for phonetic transcription is far too laborious, and requires a lot of practice to attain proficiency. It was therefore necessary to use a system that would capture the typical dialectological features of Norwegian, reflect the deviance of heritage language, yet be simple to learn and easy to use. To this end we adapted the standard described in Papazian and Helleland (2005).

3.2 Orthography

An orthographic layer was needed for two reasons. Firstly, it is required to facilitate corpus queries. Without knowledge of the phonetic system described above, any query would be hit and miss. The problem is compounded by the fact that any single word might be realised in a number of ways. Secondly, a standard orthography was required in order to carry out an automatic morphological annotation.

To meet this requirement, a semi-automatic transliterator was developed. The heart of the transliterator is a database of mappings from phonetic to orthographic representation. The database allows for a many-to-one relationship between the two, each mapping weighted according to a specific dialect. On being passed a text along with an id for the dialect at hand, the transliterator will suggest an orthographic representation. This representation can then be manually checked and corrected, before being sent back to the transliterator. At this stage, the database can be automatically trained, or adjusted, a process where new mappings are added or existing weights tweaked. This method has proved to be a reliable and effective way of producing standard orthographic annotations.

3.3 Morphology

Morphological annotation and lemmatisation was achieved by means of a TreeTagger (Schmid 1994, 1995). The tagger was trained on a version of the Oslo-Bergen tagger developed for a spoken corpus of the Oslo dialect. On the Oslo dialect material, the accuracy was measured at 96,9%, using 10-fold cross validation

(Søfteland and Nøklestad 2008). However, it cannot be assumed that the accuracy is as good for the CANS corpus. Heritage language typically contains high levels of loan words, with CANS having roughly 3%. This, along with dialectal word order, will have lowered the accuracy.

4. Querying CANS

CANS is available through the corpus search interface Glossa (Johannessen et al. 2008). The interface provides access to all three of the annotation layers mentioned above, allowing queries to contain phonetic, orthographic and morphological expressions, individually or in combination. Regular expressions can also be used on these layers. Queries can be further refined through the application of metadata constraints, the metadata being stored in a database and associated with the corresponding segments of the corpus. The layers of annotation are, of course, also available in the query result set.

4.1 Orthographic querying

By default, search queries are applied to the orthographic layer, since it does not require knowledge of the particular phonetic transcription standard adopted for this corpus. An example is given in Figure 1, which shows a search for the orthographic form of the word *var* "was/were". The first few results of this query are shown in Figure 2.

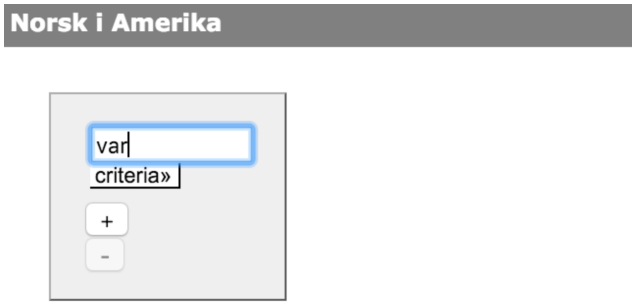

Figure 1. Querying an orthographic word form

urat # e ## e ## em i krigen # i så da **var** jeg der (uninterpretable) i Texas # når mannen min var der f

urat # e ## e ## em i krigen # i så da var jeg der (uninterpretable) i Texas # når mannen min **var** der f

han **var** i Europa # og han var e ## em

han var i Europa # og han **var** e ## em

¦ fikk telegram at han **var** ## wounded

Figure 2. Some results of the orthographic query in Figure 1

4.2 Phonetic querying

The phonetic annotation can be queried directly by selecting the appropriate options from the pull-down menu in the search interface. Figure 3 illustrates a search for the same word as in the previous example, only this time phonetically transcribed as *vaR* 'was/were', where *R* indicates the approximant [ɹ]. As shown in Figure 4, specifying a particular phonetic form yields somewhat different results from those obtained with an orthographic query; while the first three results of the two queries are the same, the last two differ.

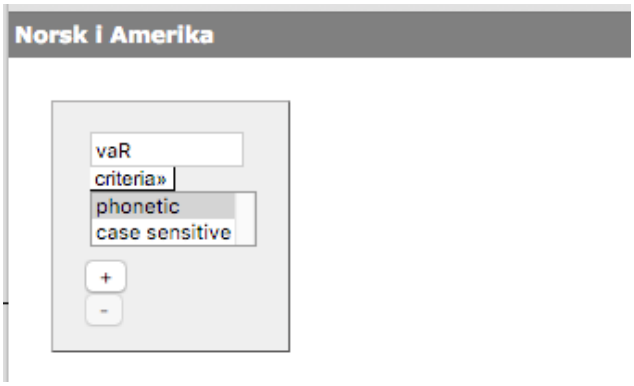

Figure 3. Querying a phonetically transcribed word

helt minn liv ja # akkurat # ee # ee # em i krigen # i så da **vaR** jei deR _ i Tekkses # nå mann minn vaR deR f-

helt minn liv ja # akkurat # ee # ee # em i krigen # i så da vaR jei deR _ i Tekkses # nå mann minn **vaR** deR f-

førn hann jekk åveR # hann **vaR** i Øråope # å hann var ee # em

åo ja # akkørat # _ nåo hu **vaR** att'n år tru _

hann **vaR** fRa # ee Flekke i Sønnfjor

Figure 4. Some results of the phonetic query in Figure 3

4.3 Morphological querying

Finally, queries can also be specified in terms of parts-of-speech and morphosyntactic features, potentially in combination with specification of orthographic or phonetic forms. Figure 5 illustrates this with a search for verbs having an orthographic form beginning with the letter "v", the results of which include those of the previous searches in addition to examples of verbs such as *venter* "wait", *veit* "know" etc.

Figure 5 Searching for verbs beginning with *v*

4.4 Multimodal result views

As a part of the first, manual stage of annotation, a basic segmentation into meaningful units was performed. Ideally, such segments are synonymous with sentences of written language. When using a tool such as Transcriber or Elan, this process of segmentation generates time codes for each segment, aligning them with the video footage of the conversation. Glossa can make use of these time codes in order to retrieve and play the sequences of video corresponding to a query result. If required, preceding and subsequent segments can also be retrieved, expanding the context of the search result. In order to facilitate reading, autocue-style highlighting is incorporated.

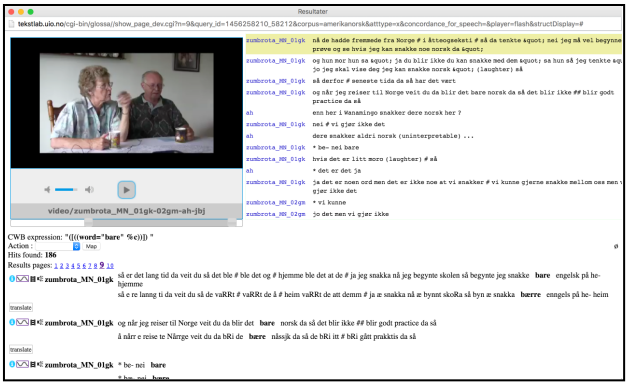

Figure 5. Query results and video

4.5 Acoustic Analysis

For a finer grained analysis of the parameters of speech, Glossa provides tools for sound visualization. Individual concordances can be selected, yielding dynamically rendered spectrograms, pitch and formants. The resulting waveforms are interactive and may be played, zoomed, filtered and exported.

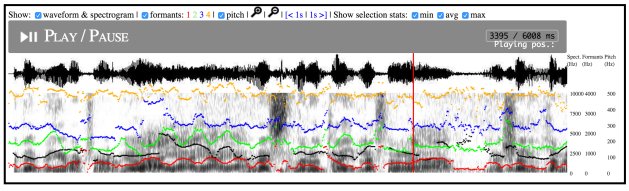

Figure 6. Waveform

References

Johannessen, Janne Bondi, Lars Nygaard, Joel Priestley, and Anders Nøklestad (2008). Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of the SixthInternational Language Resources and Evaluation (LREC'08)*. Paris: European Language Resources Association (ELRA). http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/glossa/LRECglossa_2008.pdf

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Anders Nøklestad, and Andre Lynum (2012). The Nordic Dialect Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation. European Language Resources Association*, p. 3388-3391. http://dblp.unitrier.de/db/conf/lrec/lrec2012.html

Johannessen, Janne Bondi, Øystein Alexander Vangsnes, Joel Priestley, Kristin Hagen (2014). A multilingual speech corpus of North-Germanic languages. In Raso, Tommaso; Mello, Heliana (eds.), *Spoken Corpora and Linguistic Studies*. John Benjamins Publishing Company, p. 69-83. https://www.benjamins.com/#catalog/books/scl.61.02joh/fulltext

Papazian, Eric & Botolv Helleland. 2005. *Norsk talemål*. Høyskoleforlaget, Kristiansand.

Schmid, Helmut (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing,* Manchester, UK

Schmid, Helmut (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Søfteland, Åshild and Nøklestad, Anders. 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. In Johannessen, Janne Bondi and Kristin Hagen (eds.), *Språk i Oslo. Ny forskning omkring talespråk*. Novus, Oslo.

Web Sites

Elan: tla.mpi.nl/tools/tla-tools/elan/
Glossa corpus search tool: https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/glossa.html
Oslo-Bergen Tagger: tekstlab.uio.no/obtny/english/
Text Laboratory: www.hf.uio.no/iln/english/about/organization/text-laboratory/
Transcriber: trans.sourceforge.net/en/presentation.php
TreeTagger: www.ims.unistuttgart.de/projekte/corplex/TreeTagger

PHYSIOSTRESS: A Multimodal Corpus of Data on Acute Stress and Physiological Activation

Patrice Boucher[1], Pierrich Plusquellec[2], Pierre Dufour[2], Najim Dehak[3], Patrick Cardinal[1], Pierre Dumouchel[1]

[1]École de technologies supérieures,
1100, rue Notre-Dame Ouest, Montréal (Qc). H3C 1K3,
[2]Centre for Studies on Human Stress, Centre de Recherche de l'Institut Universitaire
en Santé Mentale de Montréaland École de Psychoéducation, Université de Montréal,
7401, rue Hochelaga. Montréal (Québec) H1N 3M5
[3] Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139, United States
patrice.boucher.1@etsmtl.net, pierrich.plusquellec@umontreal.ca, ipierredufour@hotmail.com
najim@csail.mit.edu, patrick.cardinal@etsmtl.ca, pierre.dumouchel@etsmtl.ca

Abstract

This paper presents PHYSIOSTRESS, our data corpus of acute stress and physiological activation. It includes 26 experiments of acute stress based on the Trier Social Stress Test, and 24 experiments which combine a social task and a physical activity. These experiments were accomplished by 13 men and 13 women between 20 and 49 years old without identified cardiac disease, respiratory disease neither mental disease. The psychosocial situation of each participant was evaluated from 5 questionnaires including the *Rosenberg Self-Esteem Scale*, the *Perceived Stress Scale*, the *Trier Inventory for the assessment of Chronic Stress*, the *International Physical Activity Questionnaire* and a socio-demographic questionnaire. During each experiment, we record a derivation of ECG, respiratory data, the momentum of the subject (from accelerometers), the internal sounds of the body, audio and videos of the subjects and the 3D movements of the subjects (from a Kinect Microsoft sensor). The level of stress of each subject (no stress, low stress, medium stress or high stress) is annotated according to three references including: the stress felt by the subject, the stress apparent (annotated by two observers) and the subject level of salivary cortisol. PHYSIOSTRESS is a rare corpus of acute stress which combines measurements of heart and respiration with annotations of the salivary cortisol level, namely a standard in medical research in the evaluation of acute stress.

Keywords: Acute Stress, Corpus, Heart Rate Variability, Respiration, Audio, Features, Cortisol

1. Introduction

Chronic stress is a leading factor of numerous mental and physical disorders such as professional burnout, depression, addiction, anxiety, obesity, cancer, immunity diseases and cardiac diseases (Maddock and Pariante, 2001). This is, consequently, essential to rapidly detect individuals who experience repeated acute stress to intervene before they develop those disorders. In literature, relevant studies have shown a low correlation between the stress felt by individuals and their physiological stress (Lupis et al., 2014). This is thus essential to pursue development efforts in novel physiological tools for measuring the level of stress of individuals. This would enable efficient interventions through individuals who suffer from repeated stress, whether they be patients, specialists or workers.

The effect of stress on secretion of cortisol has been observed during the time of the introduction of the concept by Selye (1950). Many of his contemporary studied how stressful experiments influence the level of cortisol on animals and human (Mason, 1968). In clinical research, the evolution of the level of acute stress is commonly assessed based on the level of cortisol in saliva (for short-time studies) and in hair (for long-time studies). This manner to evaluate the stress has the advantage to be consistent with the initial concept of stress as defined as a response to a psychological threat to the body homeostasis. This answer has been decomposed through a hormonal reaction from the hypothalamus to the adrenal glands with secretion of cortisol and other hormones (Allen et al., 2014). However, evaluating short time stress by measurements of the salivary cortisol imposes noticeable limitations to the researches in costs, logistic, complexity of interpretation and sampling frequency.

These limitations encourage the development of other solutions to monitor acute stress. Some recent efforts aim at developing wearable, non-intrusive, devices that could real-time monitor acute stress from ubiquitous sensors of heart, respiration, skin and brain (Healey and Picard, 2005; Benoit et al., 2009; Chanel et al., 2009; Setz et al., 2010; Rigas et al., 2011; De Santos Sierra et al., 2011; Kumar et al., 2012). All of these systems require to establish a prediction model able to draw the relationship between the physiological inputs (e.g. from the heart activity, respiration, videos) and a correspondent level of stress. In order to establish this relationship, we need a solid corpus of data in which the physiological inputs of various subjects are recorded while the level of stress is annotated in the course of relevant experiments in which each subject encounters various levels of stress and various psycho-physiological states.

Previous data corpus on acute stress record physiological data (e.g. heart activity, respiration) with rough technic of annotations of the level of stress. Typically, the level of stress is annotated according to the context (Chanel et al., 2009; De Santos Sierra et al., 2011). In this view, the subject is supposed to be stressed when he does the selected

stressful task, and not stressed otherwise. This approach involves a yes/no stress level. In some cases, the stress is rather assessed by an observer (Healey and Picard, 2005), and in other cases by the subject himself (Healey and Picard, 2005; Rigas et al., 2011; Kumar et al., 2012). Unfortunately, previous corpus has not yet considered a more objective and precise index of stress, the variation of salivary cortisol levels, which is an inescapable reference for researchers in the field of stress studies. As a consequence, the devices build from these corpus could at best predict a rough level of stress, with an accuracy and a certitude that do not meet the standard for researchers in stress.

In our work and thanks to an interdisciplinary team of researchers, our objective is to build a corpus of data which could serve in development of reliable stress monitoring devices for researches in stress and for public health. In this way, we carefully studied the state-of-art standards in stress researches in order to (1) choose an efficient stress task; (2) choose reliable references of stress for annotations; and (3) best control any bias that could result in the procedure.

For the first point, we have chosen a widely employed procedure to induce acute stress in current research, namely the TSST: Trier Social Stress Test (Kirschbaum et al., 1993). Many authors have demonstrated the impact of the TSST through various systems of the human body including the respiratory system, the nervous system, the endocrine system, the cardiovascular system and the lymphatic system (Kudielka et al., 2007).

For the second point, we selected three references of stress: the stress felt by the participant, the apparent stress evaluated by two observers and the variation of the salivary cortisol level across the TSST. While the last reference of stress (with the cortisol) appears as the more reliable, the two others (stress felt and apparent stress) are still very relevant since they can be evaluated more frequently. Moreover, those annotations can be very interesting for researchers interested in studying or modeling relationship between them. We expect that these three references could be combined in a single annotation, producing a more accurate reference of stress along the experiment.

For the last point, we have mainly ensure that the data includes experiments which cause similar physical activities for different level of stress. This point is important to force a stress modeling that exploits specific features of stress without being biased by the physical activity. Hence, we add an experiment day which attempts to reproduce the physical activity of the first day of TSST, but this time without stress.

2. Physiological measurements

We employed sensors to record: (i) the cardiac activity; (ii) the respiratory activity ; (iii) the body movements; (iv) the facial and body expression; and (v) the voice of the participant. Standard cameras record the facial expression and the body expression, while a standard microphone records the participant's voice during the stress task, the social task and the physical task. The body movements are tracked by the Kinect II Microsoft sensor.

Audio features are extracted with OpenEars (Politepix, 2016). The videos are employed by two observers who annotate the apparent stress of the subject; and by *Facet* (iMotions Biometric Research Platform, 2016) to add extra information to the corpus related to the emotions of the participant. We plan to extract movements of the skeleton from the Kinect records, as described in Arai and Asmara (2013). This information could be used by researchers who are interested in the relationship between the body language, stress and emotions. We plan to publish all anonymous information on *Physionet* (Physionet, 2016). The wearable sensors are detailed below.

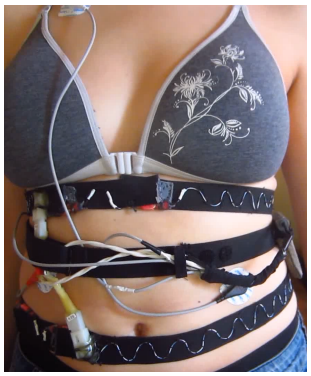

Figure 1: The Hexoskin recorder is linked directly on the skin with disposable electrodes and with customs respiratory belts.

2.1. Adapted Hexoskin Sensor

We employed the wearable biosensor Hexoskin from Carre Technologies, which provides real-time data from the heart, the respiration and the motions including:

- A derivation of electrocardiogram (ECG);
- Movements of the abdomen and the thorax, from which we can extract respiratory data (this technic is known as respiratory inductance plethysmography);
- 3D accelerations on axis x, y and z, from which we can compute the momentum of the body.

This sensor includes a record device that connects to a wearable shirt. It is specially conceived for a sports usage, which supposes a minimum level of transpiration to ensure a good electric contact between the shirt and the skin. In our testing phase of the device, before all experiments, we found that the signal was quite noisy when the devise was used without a significant level of physical activity. In order to overcome this issue, we use electrical wires to plug the record device directly on the skin through disposable electrodes; and link the wires of respiration to two adjustable belts on the thorax and the abdomen (see Fig. 1).

The corpus provides raw measurements of those sensors as well as more abstract features including cardiac intervals, the momentum and respiratory features (respiration rhythm, inhalation time, expiration time, inhalation magnitude, expiration magnitude). Cardiac intervals and respiratory features are extracted from a custom program. Cardiac intervals are then manually inspected and corrected by a

human who is very familiar with ECG. This ensures a good accuracy of this feature despite some noisy sequences.

3. Psychosocial Measurements

The participants must fill 5 questionnaires: (1) the *Rosenberg Self-Esteem Scale* (Robins et al., 2001), (2) the Perceived Stress Scale (Cohen, 1988), (3) the Trier Inventory for the assessment of Chronic Stress (Schulz and Schlotz, 1999), (4) the International Physical Activity Questionnaire (Craig et al., 2003) and (5) a 9-item socio-demographic questionnaire. Self-esteem influences psychosocial stress, namely an important stressor for the *Trier Social Stress Test* presented in Section 4.. The second questionnaire measures the stress felt by the participant in his environment and his perception of his capacity to front stressful situations. The third questionnaire measures the level of chronic stress of the participants, which could influence their stress felt and their level of cortisol. The fourth questionnaire is used to explain differences, among the participants, in the cardiovascular response during our task of physical activity. The last questionnaire allows us to account for different factors that could influence results, namely: the age of the participant, his ethnic origin, his socioeconomic status and his education level.

4. Description of the Experiments

The experiments were conducted at the Center for Studies on Human Stress, in Montreal (Canada). We recruited participants through classified advertising sites covering Montreal as well as in bulletin boards (in universities, hospitals, stores, sports centers) and in social networks. Any volunteers were free to apply. However, we accepted only those who respects the following criteria of inclusion: has no identified cardiac, respiratory or mental disorders; has never done the stress task (Trier Social Stress Test); does not consume drugs, contraceptives or hormones (which ones could affect the level of salivary cortisol); is not pregnant.

4.1. Day 1: Trier Social Test Task (TSST)

At the arrival, an assistant explained the procedure to the participant, presented him a consent form and proposed him to sign it. The assistant then explained technical details to the participant (about cortisol sampling and annotation of stress felt), installed all sensors and asked to the participant to fill the 5 questionnaires (see Section 3.).

TSST is considered as a reference test in psychoneuroendocrinological studies (Kirschbaum et al., 1993; Kudielka et al., 2007) to produce moderate stress on a majority of participants. It is subdivided in three phases: (i) a period of waiting in a room which serves as a baseline for the cortisol level (no stress is induced in this phase); (ii) a phase during which the participant is asked to do, in a second room, an oral presentation and an arithmetical task in front of three evaluators; (iii) a last phase of recovery in which the participant waits in the initial room.

During the second phase, the participant has for instruction to convince three evaluators (two in our version) that he is the best candidate for a monitor job in a summer camp with children. The participant is advised that the presentation will be recorded by cameras and microphones. The participant has 5 minutes to prepare the presentation in front of the evaluators (in our version: 10 minutes in the first room), 5 minutes to do it and 5 minutes to make an arithmetical task. The evaluators, men and women, strive to stay neutral, without emotional expression.

A single camera records the periods of waiting before and after the stress task. The stress task is recorded by two cameras for the head and the body expression, a 3D Kinect sensor and a microphone.

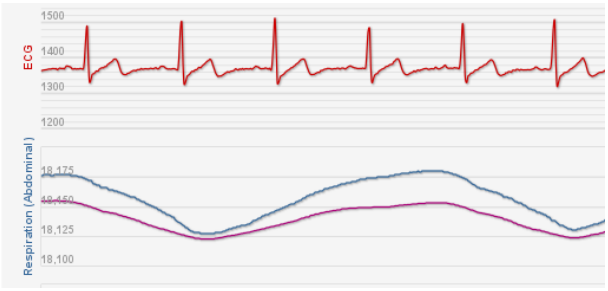

Figure 2: Example of normal ECG and respiratory signal.

4.2. Day 2: Social Task and Physical Activity

The second day is built in such a way that the participant reproduces the physical activity of the first day, but without stress. Hence, the stress task was replaced by a 10 minutes social task in which the participant talked freely of anything with the assistant, who strives to be empathic, friendly and relaxed to comfort the participant (as in a casual talk show). Just after, the assistant guided the participant for 10 minutes series of 30 seconds physical exercises (stand-still running, squats) each broken by 10 seconds of rest. The two periods of waiting and the sensor setup were the same as for the first day.

5. Results

We met 30 volunteers from June 2015 up to January 2016, including 14 women and 16 men between 20 and 49 years old. One man and one woman does not complete the second day and one man dismiss during the first day. In summary, we have 26 valid experiments of TSST and 24 valid two-day experiments.

The missed records are caused by various incidents including in particular: a damaged electrical wire or connection, a badly connected recorder or wire, a discharge of the recorder or electrodes that peel off the skin.

Figure 2 shows a record of the adapted Hexoskin device, with the ECG at top, the abdominal and thoracic expansion in the middle and the 3-axis accelerations at the bottom. We can observe noise, sometimes significant, in the ECG signal in particular when the subjects are in physical activity. This noise comes from muscular and respiratory artefacts as well as from static electricity produced by the shirt friction on the sensors. The noisy ECGs were carefully inspected in order to assess a good beat detection required to compute cardiac intervals. Some small sequences, related to other problems with the recording setup, are still unusable. These sequences total about 15 minutes on 84.5 hours of

record time. In summary, 99.7% of the ECGs provided in the experiments are usable. The audio, videos and the 3D records from the Kinect have not yet been analyzed.

6. Conclusion

We have presented in this paper our corpus on acute stress, which combines multi-source records of electrocardiogram, respiratory data, movements, videos and audios of subjects who experimented a stressful task as well as a non-stressful social task and a physical activity. These records represent 84.5 total hours annotated with three references of the acute stress including the stress apparent, the stress felt and the stress measured by the level of cortisol. The corpus will also provide extra information about the emotions of each participant as predicted by the Facet software (iMotions Biometric Research Platform, 2016) based on the decoding of facial expression. We aim at publishing all anonymous data of the corpus in *Physionet* (Physionet, 2016) for researchers interested in studying the physiological activity related to different stress and affect states, and those interested in stress modeling, emotion modeling, and non-verbal modeling. We expect that our corpus will contribute to enabling the design of state-of-the-art devices able to monitor acute stress in stress research. Such a device could subsequently serve public health in stress management programs.

7. Acknowledgements

The authors acknowledge all financial support from the *Fonds de recherche Nature et technologies* (grant 176847), Pierre Dumouchel and Pierrich Plusquellec as well as each member of the Center for Studies on Human Stress who contributes to the corpus realization, namely Jeanne Bettez, Charles Boisvert, Florence Landry, Yannick Fouda, Amélie Paulus, Héliéna Guillet and Nadia Durand, Nathalie Wan and Johanne Beauséjour. The authors also acknowledge Alexis Martin and Jérémie Voix for their ear device.

8. Bibliographical References

Allen, A. P., Kennedy, P. J., Cryan, J. F., Dinan, T. G., and Clarke, G. (2014). Biological and psychological markers of stress in humans: Focus on the trier social stress test. *Neuroscience & Biobehavioral Reviews*, 38:94–124.

Benoit, A., Bonnaud, L., Caplier, A., Ngo, P., Lawson, L., Trevisan, D. G., Levacic, V., Mancas, C., and Chanel, G. (2009). Multimodal focus attention and stress detection and feedback in an augmented driver simulator. *Personal and Ubiquitous Computing*, 13(1):33–41.

Chanel, G., Kierkels, J. J., Soleymani, M., and Pun, T. (2009). Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607 – 627.

Cohen, S. (1988). Perceived stress in a probability sample of the united states.

Craig, C. L., Marshall, A. L., Sjostrom, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J. F., and Oja, P. (2003). International physical activity questionnaire: 12-country reliability and validity. *Med sci sports Exerc*, 195(9131/03):3508–1381.

De Santos Sierra, A., Sanchez Avila, C., Casanova, J. G., and del Pozo, G. B. (2011). A stress-detection system based on physiological signals and fuzzy logic. *Industrial Electronics, IEEE Transactions on*, 58(10):4857–4865.

Healey, J. A. and Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):156–166.

iMotions Biometric Research Platform. (2016). Facet.

Kirschbaum, C., Pirke, K.-M., and Hellhammer, D. H. (1993). The trier social stress test–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81.

Kudielka, B. M., Hellhammer, D. H., Kirschbaum, C., Harmon-Jones, E., and Winkielman, P. (2007). Ten years of research with the trier social stress test revisited. *Social neuroscience: Integrating biological and psychological explanations of social behavior*, pages 56–83.

Kumar, M., Neubert, S., Behrendt, S., Rieger, A., Weippert, M., Stoll, N., Thurow, K., and Stoll, R. (2012). Stress monitoring based on stochastic fuzzy analysis of heartbeat intervals. *Fuzzy Systems, IEEE Transactions on*, 20(4):746–759.

Lupis, S. B., Lerman, M., and Wolf, J. M. (2014). Anger responses to psychosocial stress predict heart rate and cortisol stress responses in men but not women. *Psychoneuroendocrinology*, 49:84–95.

Maddock, C. and Pariante, C. M. (2001). How does stress affect you? an overview of stress, immunity, depression and disease. *Epidemiologia e psichiatria sociale*, 10(03):153–162.

Mason, J. W. (1968). A review of psychoendocrine research on the pituitary-adrenal cortical system. *Psychosomatic Medicine*, 30(5):576–607.

Physionet. (2016). Physiobank.

Politepix. (2016). Openears.

Rigas, G., Goletsis, Y., Bougia, P., and Fotiadis, D. I. (2011). Towards driver's state recognition on real driving conditions. *International Journal of Vehicular Technology*, 2011.

Robins, R. W., Hendin, H. M., and Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the rosenberg self-esteem scale. *Personality and social psychology bulletin*, 27(2):151–161.

Schulz, P. and Schlotz, W. (1999). The trier inventory for the assessment of chronic stress (tics): scale construction, statistical testing, and validation of the scale work overload. *Diagnostica*, 45(1):8–19.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G., and Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable eda device. *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):410–417.

A Gesture-Speech Corpus on a Tangible Interface

Dimitra Anastasiou[1], **Kirsten Bergmann**[2]
[1] Luxembourg Institute of Science and Technology (LIST)
5, avenue des Hauts Fourneaux L-4362 Luxembourg
E-mail: Dimitra.Anastasiou@list.lu

[2] Social Cognitive Systems Group, Faculty of Technology (Bielefeld University)
P.O. Box 100 131, D-33501 Bielefeld
E-mail: kirsten.bergmann@uni-bielefeld.de

Abstract

This paper presents a corpus of hand gestures and speech which is created using a tangible user interface (TUI) for collaborative problem solving tasks. We present our initial work within the European Marie Curie project GETUI (GEstures in Tangible User Interfaces). This project involves mainly creating a taxonomy of gestures used in relation to a tangible tabletop which is placed at the Luxembourg Institute of Science and Technology (LIST). A preliminary user study showed that gesturing encourages the use of rapid epistemic actions by lowering cognitive load. Ongoing corpus collection studies provide insights about the impact of gestures on learning, collaboration and cognition, while also identify cultural differences of gestures.

Keywords: human-computer interaction, pointing, tangibles, taxonomy

1. Introduction

Gesturing is a natural communication means with both inter-personal and intra-personal functions. Inter-personally, in human face-to-face interaction (HHI), co-speech gestures emphasize or supplement spoken content. Intra-personally, gestures can support cognitive processing (e.g. Ping & Goldin-Meadow, 2010), a fact which can be exploited for so-called Tangible User Interfaces (TUIs). The term TUI has been established by Ullmer & Ishii (2000) as follows: "[TUIs] give physical form to digital information, employing physical artifacts both as 'representations' and 'controls' for computational media. TUIs "[provide] tangible representations to digital information and controls, allowing users to quite literally grasp data with their hands" (Shaer & Hornecker, 2010). Kirk et al. (2009) stated that the kinesthetic memory of moving a tangible object can increase the recall of performed actions, preventing mode errors, as interacting with a physical object can be equivalent to an implicit, user-maintained mode. Likewise, Esteves et al. (2013) showed that conducting problem solving tasks on a TUI encourages rapid *epistemic* actions, and lowers cognitive load by simplifying thinking processes. Accordingly, TUIs are of particular interest for the domain of technology-enriched learning environments, especially given the option to integrate technology-based assessment (TBA). However, currently there is neither a systematic analysis of employing TUIs in the context of TBA, nor has interaction with TUIs been systematically explored in TBA yet.

In this paper we present first steps towards a detailed investigation of integrating TUIs and TBA. Our goal is to set up a data-based taxonomy of gestures used in interaction with a TUI, whereby our domain of application is collaborative problem solving as one of the most important 21st Century skills. The paper is laid out as follows: Section 2 presents related work on gesture taxonomies in general as well as the very relation of gesture and cognition as well as gesture and culture. In Section 3 we present a pre-study and its results. The design of our corpus collection studies is described in Section 4, followed by draft for the gesture taxonomy. Finally, we discuss annotation issues and draw conclusions with respect to the impact of the corpus.

2. Related Work

Gesture taxonomies have been presented in the literature both from a philological viewpoint and a human-computer interaction (HCI) viewpoint. Regarding the former, foundations about gestures were established by McNeill (1992), based on Kendon's continuum (Kendon, 1982); gestures were classified into *gesticulation*, *pantomime*, *emblem*, and *sign language*. *Gesticulation* is further classified into *iconic*, *metaphoric*, *rhythmic*, *cohesive*, and *deictic* gestures. As for the latter, from an HCI perspective, Quek (1994) created a taxonomy in HCI, classifying meaningful gestures into *communicative* and *manipulative* gestures. Manipulative gestures can occur either on the desktop in a 2-D interaction using a direct manipulation device, as a 3-D interaction involving empty-handed movements to mimic manipulations of physical objects, or by manipulating actual physical objects that map onto a virtual object in TUIs. We focus particularly on the third categorization of manipulative gestures. The most prevalent type of gesture in relation to TUIs is *pointing* or *deictic*. Moreover, Lao et al. (2009) defined *tapping*, *pressing*, and *dragging* gestures and showed that a variety of hand gestures can be constructed through these three basic movements. Karam and Schraefel (2005) made a classification of the literature about gesture interaction research (mainly user studies) since the early 1990s.

As far as hand gesture recognition is concerned, Rautaray & Anupam (2015) made a recent survey on vision-based hand gesture recognition for HCI, analysing the three main recognition techniques (detection, tracking,

recognition) as well as the required software platforms.

Gesture and cognition Alibali et al. (2000) stated that gesturing reduces the cognitive load for both adults and children, particularly during explanation tasks. Klemmer et al. (2006) pointed out that systems that constrain gestural abilities, e.g. having the hands stuck on a keyboard, are likely to hinder users' thinking and communication.

Gesture and culture Gestures and their cultural connotations have been examined, among others, by Archer (1997) and Kita (2009). Archer (1997) found that there are both cultural differences and meta-differences, i.e. more profound differences involving deeply embedded categories of meaning that make cultures unique. Kita (2009) reviewed the literature on cross-cultural variation of gesture based on four relevant factors: conventions of form-meaning association, language, spatial cognition, and pragmatics of gesture use. Moreover, we had previously defined a *locale* as a combination of language and culture as well as gesture localization as follows: *Gesture localization is the adaptation of gestures to a target locale in order to transfer the same meaning as in the original locale* (Anastasiou, 2011). The importance of addressing cultural differences in gesture use becomes more and more important in times of globalization and migration. In Luxembourg, for instance, there were 220,522 foreigners equivalent to 43.04% of the total population in 2011[1]. In this multi-lingual and -cultural society, it is essential to be aware of gestures based on other cultures, so that humans and machines communicate 'properly' by respecting other people's culture.
In our research, we aim to integrate all the aforementioned aspects by setting up and analyzing a corpus of speech-gesture use in collaborative interaction with a TUI. The goal is to create a gesture taxonomy both from a philological and HCI perspective under consideration of cognitive skills and cultural differences of gesture use.

3. Pre-Study

A preliminary user study was conducted at the Luxembourg Institute of Science and Technology (LIST) (Anastasiou et al., 2014). There were 10 groups of three people in each group; the task of the participants was to explore the relation of external parameters on the production of electricity of a windmill presented on a tangible tabletop. The goal of the study was to observe, analyze and understand the interactions between participants while collaboratively solving a task. We annotated in total 601 gestures, 334 of which were manipulative, 181 pointing, followed by 35 emblems, 28 iconic, and 23 adaptors. The gesture analysis based on this preliminary study resulted in the following gesture taxonomy:

1. ***Deictic/pointing*** gestures: point something/somewhere, such as to a(n):
    a. Object(s);
    b. TUI;
    c. First object and then TUI;
    d. Other participant(s);
    e. Collaborative pointing.
2. ***Iconic*** gestures: indicate distance, depth, or height or describe the shape of an object;
    a. *Encircling*: making a circle with fingers representing the turning of the physical object;
    b. *Moving an open hand forward/backward*: representing distance and/or asking from a participant to move the physical object forward/backward;
    c. *Moving an open hand downwards vertically*: representing depth.
3. ***Emblems***: have a direct verbal translation and can be interpreted differently by different cultures;
    a. *Holding open hand*: prompting other participants to wait or stop interaction;
    b. *Raising hand with palm up*: indicating uncertainty, questioning "what are we/you doing?";
    c. *Showing an open hand:* prompting other participants to continue interaction;
    d. *Raising finger/arm (open hand)*: indicating uncertainty, such as "I do not know";
    e. *Shaking fingers in a circular way*: indicating fuzziness, like "so and so".
4. ***Adaptors****:* are not used intentionally during a communication or interaction;
    a. Head/chin/nose scratching;
    b. Touching nose/mouth.
5. ***TUI-related/manipulative*** gestures: occur specifically in interaction with TUIs.
    a. *Placing*: taking the object from table frame and putting it on a specific position on the TUI;
    b. *Tracing*: moving the object to another place of the table by dragging it on the TUI;
    c. *Rotating*: turning the object from right to left or left to right;
    d. *Moving*: holding up the object from table and placing it somewhere else on the TU.

In general, our study showed that problem solving task on the TUI encouraged the use of rapid epistemic actions by simplifying thinking processes. This conclusion is drawn by two results of our study: (1) almost half of the gestures were not TUI-related, i.e. did not modify anything in the simulation and just helped to lower cognitive load by simplifying the thinking process (Esteves, 2013) and (2) in case of a gesture, the other participants reacted also with gestures (85.4% TUI-related gestures); this shows that modifications on the parameters could be quickly done and feedback was provided immediately. Moreover, we observed that 78,5% spoke during gesturing, which shows the tight connection between speech-gesture, as already well established in the literature (Goodwin, 1994; Ping & Goldin-Meadow, 2010).

4. Corpus collection studies

Participants To take cultural differences into account, we recruit 60 participants in our evaluation studies, separated in 3 *locales*: 20 francophone, 20 germanophone, and 20

[1] Statistics Portal: http://www.statistiques.public.lu/en/news/population/population/2012/08/20120821/index.html, 18.02.16

anglophone. The participants are minor students (15-18 years old) and recruited through public schools in Luxembourg.

Task Participants' task is based on a microworld; the three pupils will be provided with three physical objects that represent industrial facilities that produce electricity, e.g. a windmill, photovoltaics and a coal-fired power station. The objects will be given artificial names or variables in order to avoid previous knowledge. By turning the objects on the TUI (input: 0-10 scale), there are two parameters changing: i) the electricity generation and ii) CO2 emission. The pictures depicted on the tabletop will be accordingly adapted to the output values. This task is similar to tasks given in the international large-scale educational Programme for International Student Assessment (PISA) programme.

Study Setup The TUI employed for the study at LIST institute in Luxembourg is realised as a tangible tabletop (75x120 cm). Physical objects can be manipulated on the table in order to explore different factors. The table provides visual feedback in real-time and displays the effects with pictures and animations.

Data A multimodal corpus of video volume≈9 h and 200GB is currently collected. The Kinect 2.0[2] depth sense camera is used for recognition of the spatial position of the participants, proximity (between users and between users and tabletop) and their gestures. The light-weight and extensible software framework TULIP (Tobias et al. 2015) is used, which combines the TUI interaction paradigm and software engineering principles. We will draw upon the collaborative problem solving assessment approach that was employed in PISA 2015. We will follow the MicroDYN framework of Greiff et al. (2012), a new approach for computer-based assessment of CPS based on linear structural equations. This methodology allows to formally describe everyday activities by means of variables, outcomes and their interconnectedness.

4.1 Annotation challenges

Our user studies will result in a multimodal corpus of speech and gesture that will be annotated with ELAN (Wittenburg et al. 2006) and NEUROGES (Lausberg & Sloetjes, 2015). It will be examined whether speakers of typologically different languages exhibit differences in their gestural patterns, how gestures are coordinated with intonation and to which degree are semantically and pragmatically co-expressive with the verbal utterance. Moreover, we will examine which part-of-speech (PoS) users used in every gesture phase of the gesture unit: *preparation, stroke,* and *retraction.* (Kipp, 2004). For instance, for the sentence "This belongs here", they might use the pointing gesture synchronously with the word *this* or the word *here* or they might use two subsequent gestures. Our gesture taxonomy will include *locale*-specific gestures, as participants from three different locales are recruited.
In addition, in our evaluation studies it is examined whether the spatial context/proximity affects the gestural performance. For example, we observe whether it plays a role where exactly the participant stands in relation to the physical objects or the objects in relation to the tabletop. Through these dimensions, we assess collaborative complex problem solving and reasoning skills.

4.2 Gesture taxonomy

Our gesture taxonomy is partially based on our draft taxonomy (see Sec. 3) with the difference of putting together *pointing* and *iconic* gestures under *physical,* and *emblems* and *adaptors* under *affective* gestures. Moreover, this taxonomy is extended by adding the category of *collaborative* gestures. Because of page constraints, only a few layers are presented here.

1. *Physical*
    a. *Pointing*
        i. *Single-handed*
            - *Object*
            - *TUI, etc.*
        ii. *Bimanual*
    b. *Iconic*
2. *Manipulative*
    a. *Placing*
    b. *Removing*
    c. *Rotating*
    d. *Tracing*
3. *Affective*
    a. *Emblems*
    b. *Adaptors*
4. *Collaborative*
    a. *Symmetry*
        i. *Symmetric*
        ii. *Asymmetric*
        iii. *Partially symmetric*
    b. *Parallelism*
    c. *Additivity*

The last category is cross sectoral, as it combines gestures from other categories. Collaborative gestures have been examined in the literature by Block et al. (2015), Tang et al. (2006), and Morris (2006). Morris (2006) stated that symmetry axis refers to whether participants in a cooperative gesture perform identical actions or distinct actions, while parallelism is defined as the relative timing of each contributor's axis. An *additive* gesture is one which is meaningful when performed by a single user, but whose meaning is amplified when simultaneously performed by all members of the group.
Very often in the literature about gesture taxonomies in HCI, gestures are mixed with verbal utterances. In our annotation scheme, speech is considered as a separate modality and will be first transcribed, then annotated based on Conversational Analysis and third examined in temporal coordination with gestures. Verbal utterances are categorized into substantial or pragmatic (Kendon, 2004), interpretation, conflicts, negotiation, etc. We plan to extend the CPS model of dialogue by Blyloke and Allen (2005) and Hmelo-Silver (2003).

5. Conclusion

The research presented here addresses a practical application field of HCI and Interaction Design: TUIs. In the literature it has been shown that gesturing can lower cognitive load, a fact that we also substantiated in our pre-study. The main objective of our current research is to

[2] http://www.xbox.com/en-US/xbox-one/accessories/kinect-for-xbox-one

explore through user studies the gestural performance of users while interacting on a TUI in a collaborative problem solving task and in addition, what kind of effect does this the gestural performance have on 21th Century skills. We explore these aspects through a corpus collection study with 60 pupils from three different locales. The data will be analysed with respect to how participants interact with each other (gestures from HHI perspective) and with objects/TUI (HCI perspective) and how these both kinds of interaction facilitate the technology-based assessement.

As far as the impact of corpus is concerned, both gesture and speech researchers will benefit from its existence and annotation. Moreover, educators and pupils will learn not only about power grids, but at a more abstract level, about computer-mediated collaborative problem solving in general. At a higher level, we will provide guidelines for future applicability of TUIs in PISA. Moreover, as we will have 3D data as output of the Kinect, gesture recognition researchers can train their systems and increase the accuracy, particularly for hand and finger gestures which is a quite big recognition challenge. Particularly, such a scenario is particularly challenging, as there are many users crossing in front of each others or placing hands on the top of other hand(s).

6. Acknowledgements

This research is funded by the Marie Curie Individual Fellowship project *GEstures in Tangible User Interfaces*.

7. Bibliographical References

Alibali, M.W., Kita S, Young A. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language & Cognitive Processes*, 15, pp. 593--613.

Anastasiou, D., Maquil, V., Ras, E. (2014). Gesture Analysis in a Case Study with a Tangible User Interface for Collaborative Problem Solving. *Journal on Multimodal User Interfaces*, Springer.

Anastasiou, D., (2011). Speech Recognition, Machine Translation and Gesture Localisation. *TRALOGY: Translation Careers and Technologies: Convergence Points for the Future,* 3 - 4 March, Paris, France.

Archer, D. (1997). Unspoken diversity: cultural differences in gestures. *"Visual Sociology" Qualitative Sociology*, 20(1), pp. 79--105.

Block, F. et al. (2015). Fluid Grouping: Quantifying group engagement around interactive tabletop exhibits in the wild. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 867--876.

Blyloke, N., Allen, J., (2005). A collaborative problem-solving model of dialogue. *Proceedings of the SIGdial Workshop on Discourse and Dialog*.

Esteves, A. et al. (2013). Physical games or digital games?: Comparing support for mental projection in tangible and virtual representations of a problem solving task. *Proceedings of TEI*, pp. 167--174.

Goodwin, C. (1994). *Professional vision*. American Anthropologist, 96(3), pp. 606--633.

Greiff, S., Wüstenberg, S., Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement*, 36, pp. 189--213.

Hmelo-Silver, C.E. (2003). Analyzing collaborative knowledge construction: multiple methods for integrated understanding. *Computers & Education* 41, pp. 397--420.

Karam, M., Schraefel, mc(2005). *A Taxonomy of Gestures in Human Computer Interactions,* Technical Report.

Kendon, A. (1982). The study of gesture: some observations on its history. *Rech Semiot Semiot Inq* 2(1), pp. 25--62.

Kendon, A. (2004). *Gesture-Visible Action As Utterance*, UK: Cambridge University Press.

Kirk, DS. et al. (2009). Putting the physical into the digital: issues in designing hybrid interactive surfaces. *Proceedings of BCS HCI 2009*, pp 35--54.

Kipp, M. (2004). *Gesture generation by imitation—from human behavior to computer character animatio*n. PhD Dissertation, Boca Raton, Florida.

Kita, S. (2009) Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, 24(2), pp. 145--167.

Klemmer, S.R., Hartmann, B., Takayama, L. (2006). How bodies matter: Five themes for interaction design. *Proceedings of DIS 2006 Conference on Designing Interactive Systems*, pp. 140--149.

Lao, S. et al. (2009). A gestural interaction design model for multi-touch displays. *Proceedings of the British HCI-Group*.

Lausberg, H., Sloetjes, H., (2015). The revised NEUROGES-ELAN system: An objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture. *Behavior Research Methods*.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

Morris, M.R. et al. (2006). Cooperative gestures: multi-user gestural interactions for co-located groupware. *Proceedings of CHI 2006*.

Ping, R., Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science,* 34, pp. 602--619.

Quek, F. (1994). Toward a vision-based hand gesture interface. *Proceedings of the virtual reality, software and technology conference*, pp. 17--31.

Rautaray, S., Anupam A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review,* 43(1), pp. 1--54.

Shaer, O., Hornecker, E. (2010). Tangible user interfaces: past, present and future directions. *Found Trends Hum Comput Interact*, 3 (1–2), pp.1--137.

Tang, A. et al. (2006). Collaborative coupling over Tabletop Displays. *Proceedings of CHI 2006.*

Ullmer, B, Ishii, H. (2000) Emerging frameworks for tangible user interfaces. *IBM Systems Journal*, 39, pp. 915--931.

Wittenburg, P. et al. (2006). ELAN: a professional framework for multimodality research. *Proceedings of the 5th Conference on Language Resources and Evaluation*, pp. 1556--1559.

Filled pauses, Fillers and Familiarity in Spontaneous Conversations

Costanza Navarretta
University of Copenhagen
Njalsgade 140
costanza@hum.ku.dk

Abstract

This paper presents a pilot study of the use of fillers, filled pauses and co-occurring gestures in Danish spontaneous conversations between two or three people who know each other well and in dyadic first encounters. Fillers, such as the English *uh* and *um* are very common in spoken language and previous research has indicated that they function as interaction management and/or discourse planning signals. In particular, filled pauses have been found to be very frequent when speakers have to express something challanging. Previous research has also found a correlation between the degree of familiarity of the conversants and the frequency of speech overlaps and feedback unimodal and multimodal signals. In this study, we investigate whether the frequency of fillers and filled pauses and the familiarity degree of the conversants are connected and we hypothesize that fillers and filled pauses will occur more frequently in the first encounters than in conversations between persons who know each other because the communicative setting is more challenging in the former case. The results of our study confirm this hupothesis. our study also shows that fillers and filled pauses co-occur with gestures more frequently in the first encounters than in the conversations between acquainted participants. This might be due to the fact that people who are familiar do not need to signal that they want to take or give the turn as strongly as people who do not know each other. However, since many factors can influence communication, these findings should be confirmed on more data.

Keywords: Fillers, multimodal Corpora, Gestures

1. Introduction

This paper is about the use of fillers and filled pauses in different types of Danish conversation and their possible relation to the degree of familiarity of the conversants. Fillers, such as the English *uh* and *um*, are common in spoken language and are often accompanied by pauses (filled pauses) and can co-occur with more types of gesture. Language specific studies of fillers have focused on their types and uses in various languages and contexts. For example, researchers have reported that the vocal-nasal filler *um* in English occur most frequently in the beginning of sentences or larger discourse segments signaling discourse planning, while the vocal *uh* and *ah* often precede a content word indicating lexical retrieval (Shriberg, 1994; Clark and Tree, 2002). A comparative study of fillers in Dutch, English and German (de Leeuw, 2007) indicates language specific differences in the contexts in which vocal and vocal-nasal fillers occur.

Fillers and filled pauses have many functions which are not mutually exclusive (Clark and Tree, 2002). More specifically, they are connected to interaction management and can signal feedback giving, feedback eliciting (Allwood et al., 1992; Allwood, 2001), and turn management (Maclay and Osgood, 1959; Duncan and Fiske, 1977; Clark and Tree, 2002). Moreover, they can signal cognitive processes related to the planning of discourse. Reynolds and Paivio (Reynolds and Paivio, 1968) report,for example, that students used more pauses and filled pauses when they had to define abstract objects than concrete objects. These findings have also been confirmed by Rochester (Rochester, 1973) who reports that speakers used many filled pauses when they had to choose between more options or had to express complex or in other ways challenging content. Finally, filled pauses as well as other types of pause can signal lexical retrieval (Krauss et al., 2000).

Fillers and filled pauses can also have positive effect on the listener. For example, fillers can help the listener to respond more quickly to an instruction in cases where the speaker repairs a preceding error (Brennan and Schober, 2001) and fillers can signal to the listener that the speaker is going to refer to a less accessible referent (Arnold et al., 2007; Barr and Seyfeddinipur, 2010).

Fraundorf and Watson (Fraundorf and Watson, 2011) found that filled pauses had a positive effect on the listener's late recall of complex discourse, and they did not notice the same effect when they replaced the fillers with coughs of the same length. Finally, software agents have been judged to be more human-like when they used fillers (Cassell et al., 1994; Traum and Rickel, 2002; Pfeifer and Bickmore, 2009).

Since face-to-face communication is multimodal, fillers and filled pauses also co-occur with gestures. Research on English monologues has shown that speakers use fewer filled pauses when they produce more hand gestures and vice-versa (Christenfeld et al., 1991). Rausher and colleagues (Rauscher et al., 1996) find that speakers use more filled pauses while describing spatial content if they are not allowed to gesture. Speech is not affected significantly by the non-gesture condition in their study if the speakers describe non spatial content. Finally, Esposito et al. (Esposito et al., 2001) report that filled pauses in English often co-occur with gestural holds and interpret these holds as having a parallel and correlated function to that of speech pauses: Gestural holds signal that the speaker is planning new gestural content just as speech pauses signal that the speaker is planning new speech content.

In a preceding study of the function of fillers , filled pauses and co-occurring gestures in Danish dyadic first encounters, we showed that the function of the gestures reinforce the function of the fillers and filled pauses (Navarretta,

2015). We also found that the most common use of the filler *mm* in the first encounters is of giving feedback and that it is often accompanied by nods, while the vocal *øh* and vocal nasal *øhm* are used similarly to the corresponding English fillers *uh* and *um*, that is *øh* often precedes a single content word, indicating lexical retrieval, while *øhm* most frequently precedes clauses, sentences, or even larger discourse units signaling discourse planning. We hypothesized that the high frequency of fillers and filled pauses in the Danish first encounters might be related to the specific communicative situation.

Since previous research has shown that increasing familiarity of the participants is related to higher occurrences of speech overlaps (Campbell, 2009) and feedback signals (Navarretta and Paggio, 2012), we are interested in determining whether degree of familiarity also influences the use of fillers. In particularwe expect that fillers and filled pauses are inverse proportional to the degree of familiarity since first encounters are a more challenging communicative situation than every day conversations between people who know each other well. We do not expect that the frequency of gestures co-occurring with fillers and filled pauses will be influenced by the degree of familiarity.

The paper is organized as follows. In section 2. we shortly describe the data used in this study and in section 3. we present and discuss the uses of fillers, filled pauses and co-occurring gestures in the conversations involving participants that know each other well and in the first encounters. Finally, in section 4., we conclude and present future work.

2. The data

The first corpus in our study consists of annotated video-recordings of dyadic and triadic spontaneous conversations involving five women aged 55+ who are family members or near friends. The conversations were recorded in private homes by researchers at the University of Southern Denmark, who then transcribed them according to conversation analysis (CA) conventions under the Danish Clarin project. Part of the data (approximately 20 minutes recordings) are included in this study. These conversations were re-transcribed with word time stamps and multimodally annotated by researchers at the University of Copenhagen under the same project. A more detailed description of these data is in (Navarretta, 2011). The participants in the conversations spoke freely while sitting around a table, drinking coffee and eating. Some of the subject addressed are family, neighbors, education and economic crisis. Figure 1 shows a snapshot from one of the dyadic conversations. Since fillers in the Danish Clarin corpus are transcribed differently than in the NOMCO corpus, we have normalized its transcriptions following the convention in the Danish lexicon *Den Danske Ordbog*[1] which was also followed in NOMCO. More specifically, *mmm*, *mmmm*, *mmmm*, *mmmmm* and so on in the Danish Clarin corpus have all been transcribed as *mm*, while fillers such as *hhh*, *hhhh*, *hhhhh* and *hhhhh* which indicate hesitations or the filler *øhm* in the Danish Clarin have been manually analyzed and classified as one of the two.

[1] http://ordnet.dk/ddo/ordbog.

Figure 1: A snapshot from a conversation between two friends

The second corpus used in the study is the NOMCO corpus of first encounters which consists of 12 dyadic conversation between 12 young students (6 males and 6 females) aged 19-32 who talked freely standing in front of each other. The encounters were audio- and video-recorded by three cameras in a studio at the University of Copenhagen. The duration of this corpus is 65 minutes. A more detailed description of the corpus is in (Paggio and Navarretta, 2011). Both corpora are multimodal annotated and the annotations use the function and shape features of gestures defined in the MUMIN annotation framework (Allwood et al., 2007). The gestures included in this study are head movements, facial expressions, body movements and head gestures. Since the granularity of the annotations is different in the two corpora, we only use the most general features in this study. These shape features are in Table 1.

Attribute	Value
HeadMovement	Nod, Jerk, HeadForward, Tilt, HeadBackward, SideTurn, Shake, Waggle, HeadOther, None
General face	Smile, Laugh, Scowl, FaceOth, None
BodyDirection	BodyForward, BodyBackward, BodyUp,BodyDown, BodySide, BodyTurn, BodyDireOther, None
Handedness	SingleHand, BothHands

Table 1: Shape features

3. Familiarity degree and fillers

In the following we present a comparison of the frequency and use of the fillers in the two corpora.

In Table 2 the total number of word occurrences and the occurrences of words per second in the two corpora are given. Pauses and sounds such as breaths and smacks which are transcribed with words in the NOMCO corpus have not been excluded in the table. The Table shows that the participants who know each other utter more words per second than the participants who meet for the first time. This might be due to the higher occurrence of speech overlaps when the familiarity degree is higher (Campbell, 2009), but this

Table 2: Filler types and their frequency in the three data types

Corpus	Words	Words/sec
Da-Clarin	5475	4.75
Nomco	13620	3.46

aspect need be investigated further. The filler types which occur in the two corpora and their frequency are in Table 3. There are 113 fillers in the Danish Clarin data and thus

Table 3: Filler types and their frequency in the three data types

Filler	Danish Clarin	Nomco
øh	31	375
mm	9	109
øhm	71	84
årh	0	9
åh	1	9
eh(m)	1	1
Total	113	587

their percentage with respect to the words is of 0.021%. In the first encounters there are 587 fillers that is 0.043% of the words. The percentage is slightly lower in the dyadic Danish Clarin conversations than in the triadic conversations (0.018 vs. 0.023), but they show the same tendency. The percentage of fillers with respect to the words is more than twice higher in the first encounters than in the conversations between people who know each other. The fact that more words per second are uttered in conversations in which there are fewer disfluencies is not surprising.

Our starting hypothesis, that there would be more fillers and filled pauses in the first encounters than in conversations between people who know each other well is thus confirmed. It must be noted, however, that other factors can influence the number of disfluencies in the two corpora, such as the age and gender of the participants, the different physical settings and the content of discourse.

In the following, we account for the gestures which co-occur with fillers in the two corpora. Table 4 shows the most common fillers and co-occurring gestures as well as the percentage of occurrences without and with gestures in the two data-sets[2]. The figures relative to NOMCO are taken from (Navarretta, 2015). The table indicates that the fillers and filled pauses can be accompanied by all types of gestures in both corpora. However, contrary to our expectations, the fillers are most often accompanied by gestures in the NOMCO corpus than in the Danish Clarin corpus. The reason for this can be the different physical setting, but it can also be related to the degree of familiarity of the participants. For example, people who know each other well do probably not need mark interaction management such as turn keeping, giving or eliciting as strongly as people who meet for the first time. Body postures are generally less frequent in the Danish Clarin data than in the NOMCO data because the participants in the former are sitting while the participants in the latter stand up.

[2]It must be noted that more gesture types can co-occur with a filler, but this is not indicated in the table.

Filler	Alone	Head	Face	Body	Hand
Da-Clarin					
øh	35%	52%	6%	1%	29%
mm	56%	22%	22%	0	0
øhm	41	37%	3%	6%	22%
Nomco					
øh	25%	50%	18%	26%	11%
mm	19%	68%	28%	19%	2%
øhm	23%	55%	38%	23%	8%

Table 4: Filler types and percentage that co-occurs gestures

A first analysis of the contexts in which the fillers occur in the Danish Clarin conversations confirms the findings in the first encounters (Navarretta, 2015). More specifically *øh* often precedes substantives, verbs, adjectives and adverbs indicating lexical retrieval, while *øhm* more often precedes a sentence signalling that a larger discourse segment is being planned. The filler *mm* is seldom used in the Danish Clarin conversations while in the first encounters it is a common feedback signal, often accompanied by nods and smiles. Whether this difference depends on the age of the participants or is related to other factors, such as regional variance should be investigated further.

4. Conclusions and Future Work

We have compared the use of fillers, filled pauses and the gestures that co-occur with them in Danish conversations between people who know each other well and in first encounters hypothesizing that people who are not familiar would use more frequently fillers and filled pauses because the communicative situation is more challenging. We did not expect differences in the frequency of gestures who co-occur with fillers and filled pauses. The results of the study confirm that the frequency of fillers and filled pauses is inverse proportional to the familiarity degree of the conversants. Not surprisingly, the participants who utter more words per second also use fewer disfluencies. The high frequency of words in the conversations between people who know each other well could be related to the high number of speech overlaps. This has not been tested but it would in line with the study by Campbell (Campbell, 2009) who found a correlation between high familiarity degree and high frequency of speech overlaps´in Japanese telephone conversations.

The analysis of the two corpora shows surprisingly that gestures co-occur with fillers and filled pauses more often in the first encounters than in the Danish Clarin conversations. This might be due to the fact that people who do not know each other must signal more explicitly through speech and gestures whether they, for example, want to keep the turn while searching for a word or giving the floor if they have difficultis in continuing talking on a subject. We did not find sifferences in the contexts in which the most common fillers occur in the two corpora, but we noticed that in the

Danish Clarin conversations the filler *mm* is not used as often as feedback signal in the first encounters. Also the reason behind this difference should be investigated further, Since the data in this study are not large and numerous factors can influence the use of fillers and filled pauses a part from familiarity degree, the results of this study should be tested on more data types and languages.

5. Acknowledgements

Thanks the researchers who collected and transcribed the Danish Clarin conversations, Johannes Wagner, Brian MacWhinney and Lone Laursen, the annotators of the Danish Nomco corpus, Anette Luff Studsgård, Sara Andersen, and Bjørn Wessel-Tolvig and last but not least my colleague Patrizia Paggio.

6. Bibliographical References

Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing. *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the LRE Journal*, 41(3–4):273–287.

Allwood, J. (2001). Dialog Coding - Function and Grammar: Göteborg Coding Schemas. *Gothenburg Papers in Theoretical Linguistics, University of Göteborg*, 85:1–67.

Arnold, J., Hudson, C. K., and Tanenhaus, M. (2007). If you say *thee uh*- you're describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33:914–913.

Barr, D. J. and Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, 25(4):441–455.

Brennan, S. E. and Schober, M. F. (2001). "how listeners compensate for disfluencies in spontaneous speech ". *Journal of Memory and Language*, 44(2):274–296.

Campbell, N. (2009). An audio-visual approach to measuring discourse synchrony in m ultimodal conversation data. In *Proceedings of Interspeech 2009*, pages 12–14.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

Christenfeld, N., Schachter, S., and Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20(1):1–10.

Clark, H. H. and Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84:73–11.

de Leeuw, E. (2007). Hesitation markers in english, german, and dutch. *Journal of Germanic Linguistics*, 19:85–114, 6.

Duncan, S. and Fiske, D. (1977). *Face-to-face interaction*. Erlbaum, Hillsdale, NJ.

Esposito, A., McCullough, K. E., and Quek, F. (2001). Disfluencies in gesture: gestural correlates to filled and unfilled speech pauses. In *Proceedings of IEEE International Workshop on Cues in Communication*, Hawai.

Fraundorf, S. and Watson, D. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of memory and language*, 65(2):161–175.

Krauss, R., Chen, Y., and Gottesman, R. F. (2000). Lexical gestures and lexical access: a process model. In D. McNeill, editor, *Language and gesture*, pages 261–283. Cambridge University Press.

Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15:19–44.

Navarretta, C. and Paggio, P. (2012). Verbal and nonverbal feedback in different types of interactions. In *Proceedings of LREC 2012*, pages 2338–2342, Istanbul Turkey, May.

Navarretta, C. (2011). Anaphora and gestures in multimodal communication. In Hendrickx, et al., editors, *Proceedings of the 8th DAARC (2011)*, pages 171–181, Faro, Portugal. Edicoes Colibri.

Navarretta, C. (2015). Fillers, filled pauses and gestures in danish first encounters. In *Abstract proceedings of 3rd European Symposium on Multimodal Communication*, pages 1–3, Dublin, September. Speech Communication Lab at Trinity College Dublin.

Paggio, P. and Navarretta, C. (2011). Head movements, facial expressions and feedback in danish first encounters interactions: A culture-specific analysis. In Constantine Stephanidis, editor, *Universal Access in Human-Computer Interaction- Users Diversity. 6th International Conference. UAHCI 2011*, number 6766 in LNCS, pages 583–690, Orlando Florida. Springer Verlag.

Pfeifer, L. and Bickmore, T. (2009). Should agents speak like, um, humans? the use of conversational fillers by virtual agents. In Z. Ruttkay, et al., editors, *Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 460–466. Springer Berlin Heidelberg.

Rauscher, F., Krauss, R., and Chen, Y. (1996). Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7:226–231.

Reynolds, A. and Paivio, A. (1968). Cognitive and emotional determinants of speech. *Canadian Journal of Psychology*, 22:164–175.

Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2:51–81.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of California, Berkeley.

Traum, D. and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of AAMAS '02*, pages 766–773, New York, NY, USA. ACM.