Proceedings from the:

3rd European Symposium on Multimodal Communication

Dublin, September 17-18, 2015

Edited by

Emer Gilmartin, Loredana Cerrato, Nick Campbell

September 7, 2016

Proceedings from the 3rd European Symposium on Multimodal Communication

Dublin, September 17-18, 2015

Edited by Emer Gilmartin, Loredana Cerrato, Nick Campbell

Linköping Electronic Conference Proceedings, No. 105
ISSN: 1650-3686
eISSN: 1650-3740
ISBN: 978-91-7685-679-6
URL: http://www.ep.liu.se/ecp/contents.asp?issue=105

Organising Committee

- Patrizia Paggio, University of Copenhagen
- Jens Allwood, University of Gothenburg
- Elisabeth Ahlsén, University of Gothenburg
- Costanza Navarretta, University of Copenhagen

Local Organising Committee

- Nick Campbell, Trinity College Dublin
- Loredana Cerrato, Trinity College Dublin
- Emer Gilmartin, Trinity College Dublin

Programme Committee

- Elisabeth Ahlsén, University of Gothenburg
- Jens Allwood, University of Gothenburg
- Kirsten Bergmann, University of Bielefield
- Nick Campbell, Trinity College Dublin
- Loredana Cerrato, Trinity College Dublin
- Alan Cienki, VU Amsterdam and Moscow State Linguistic University
- Onno Crasborn, Radboud University Nijmegen
- Mary Ellen Foster, University of Glasgow
- Marianne Gullberg, University of Lund
- Dirk Heylen, University of Twente
- Kristiina Jokinen, University of Tartu
- Stefan Kopp, University of Bielefield
- Costanza Navarretta, University of Copenhagen
- Patrizia Paggio, University of Copenhagen
- Isabella Poggi, Roma Tre
- Silvi Tenjes, University of Tartu

Preface

This volume presents a selection of papers presented at MMSYM 2015, the 3rd European and 7th Nordic Symposium on Multimodal Communication, held in Dublin, Ireland, on the 17th and 18th September 2015. MMSYM aims to provide a multidisciplinary forum, bringing together researchers from different disciplines working on multimodality in human communication and human-machine interaction. Originating in the Nordic countries, this third edition of the symposium at European level has continued to attract an international audience.

MMSYM 2015 attracted researchers whose work spans several domains, linked by the topic of multimodality. The papers are listed in alphabetical order in the table of contents, but below we briefly describe them in terms of their thematic commonalities spanning from the analysis of gestures, to the analysis of filled pauses and other interactive phenomena observed in multimodal communication. Multimodality is observed not only in communication between humans (speech communication, visual communication), but also in communication between human and machine, for instance first encounter dialogues analysed in multimodal corpora of human dialogues or experimental human-machine dialogue systems. In this volume, multimodality is seen in a wide variety of domains, including the multimodal perception of attitudes in video blogs, multimodal perception in infants with and without risk of autism, multimodality in language learners, and even multimodal aspects related to turn-taking in contemporary dance improvisation.

There are papers addressing development and learning. Lozano et al report an ongoing meta-analysis of multimodal perception in infants with and without genetic risk for autism, which they posit will shed light on the acquisition of multimodal perceptual integration during development. Two papers address gesture in adult second language learners, with Levantinou and Navaretta investigating the effects of beat and iconic gestures in aiding comprehension and recall of a second language, while Wessel-Tolvig demonstrates how the acquisition of target language gestural patterns in advanced Danish learners of Italian gives evidence of learners' shift to target language semantic representations.

Allwood and Ahlsén address the contribution of gesture and speech to the construction of meaning, proposing a framework which extends the notion of meaning potential from symbols to iconic and indexical gestures, and presenting multimodal combinations of symbols, icons, and indices in face-to-face communication. Madzlan et al investigate multimodal perception of attitude in their study on video blogs, in which they present a novel annotation scheme for attitudes and report on experiments validating their annotation scheme and investigating how the different modalities jointly and separately contribute to perception of attitudes.

Several papers investigate first encounter dialogues, reporting analyses of multimodal corpora of human dialogues or experimental human-machine dialogue systems. In their respective papers, Navaretta and Paggio both address the interplay of multimodal elements of first encounter dialogues in the Danish language NOMCO corpus. Paggio reports on an analysis of temporal alignment between head movements and associated speech segments, while Navaretta investigates fillers, filled pauses, and co-occurring gestures in terms of their function, and contrasts her findings for Danish with previous work on other languages. Jokinen investigates automatic detection of co-speech gesturing in first encounter conversations, focussing on a top-down bottom-up paradigm combining human annotation and automatic analysis of video data, and discussing the applications of such technology to automatic dialogue systems. Ólafsson et al focus on the very first stages of interaction with strangers, outlining their Explicit Announcement of Presence (EAP) model, and reporting on a qualitative study of video recordings of humans approaching strangers to ask for directions, the design of a multimodal virtual agent incorporating this functionality, and a pilot user study of the system in the context of aiding second language acquisition in Icelandic.

A different kind of turntaking, that found in contemporary dance improvisation, is investigated in Evola et al's contribution, which describes the collection and annotation of recordings of improvised dance performed by experts, a micro analysis of turntaking between performers based on bodily movements and gaze, and a macro-analysis comparing the data with analogous data from non-performers.

Anastasiou et al and Cummins and Byrne treat the establishment of awareness and co-presence in communication in their respective contributions. Anastasiou et al present a Wizard-of-Oz study on awareness signals, where a smart object such as a lamp is used to demonstrate the potential for communication with a distant colleague before verbal or written communication across a network begins. Cummins and Byrne investigate the establishment of co-presence, proposing that the technical requirement for this across network relies on the establishment of zero-mean lag in communication. They discuss different ways of thinking about this problem and outline possible routes to this goal.

Brueck analyses multimodal representation of shared cultural knowledge pertaining to spatial orientation and conceptualisation in Kreol Seselwa, a French creole spoken in the Seychelles. In the study, she investigates the contribution of voice, gesture, and cultural factors including geography to speakers' use of frames of reference in spatial reference.

The range of work reflected in the papers presented here demonstrates depth and breadth of current research into multimodality and reflects the high level of interest from several disciplines in questions of how best to analyse the full range of signals and cues present in various types of interaction. We hope that this collection will excite further interest in the field, help maintain the momentum of multimodal studies, and contribute to the continuing success of the MMSYM symposia.

E. Gilmartin, L. Cerrato, N. Campbell August 2016

Table of Contents

Meaning Potentials in Words and Gestures

Jens Allwood [1], *Elisabeth Ahlsén*[1]

[1] SCCIIL Interdisciplinary Center, University of Gothenburg
jens@ling.gu.se, eliza@ling.gu.se

Abstract

This paper addresses the question of what and how gestures and speech, respectively, contribute to the construction of meaning. A point of departure of is the notion of "meaning potential" which we apply to both unimodal gestures and unimodal vocal-verbal units, as well as to multimodal vocal-gestural units, [1]. The purpose of this paper is to explore the notion of "meaning potential", not only for speech, but also for gesture. Specifically, we want to discuss the possibilities of extending the notion of a meaning potential for a symbolic sign (e.g. a word) to iconic and indexical signs.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Purpose

The purpose of this paper is to explore the notion of "meaning potential", not only for speech, but also for gesture. Specifically, we want to discuss the possibilities of extending the notion of a meaning potential for a symbolic sign (e.g. a word) to iconic and indexical signs. The reason for this is that the non-verbal gestures accompanying speech (co-verbal gestures) are iconic and indexical. So if the notion of meaning potential can be used also in relation to such gestures, a significant step in providing an account of meaning in multimodal communication will have been taken.

The theory we will present will thus be part of a sketch of some of the steps towards a cognitive semiotic theory of the semantics/pragmatics of multimodal communication

2. Background and Points of departure

Below we will briefly introduce some notions our analysis is based on, namely: the notions of communication and multimodal communication and the Peircean three modes of activating information (index, icon ad symbol) [2].

2.1. Communication and Multimodal communication

The notion of communication we will be presupposing is a notion where communication is seen as productive activation and receptive co-activation of shared content (information/understanding), while drawing on contextual resources.

In multimodal face-to-face communication, this means that a speaker produces speech and gestures to be shared with a listener in a process involving co-activation of the produced content. It is important to note that there is a mutual flow of information between speaker and listeners, so that a speaker not only speaks and gestures, but also perceives and understands his own communication as well as simultaneous words and gestures from co-communicators. Similarly, a listener not only perceives and understands, but also behaviorally reacts, for example, by verbal and gestural feedback.

Both speakers and listeners make use of the context in which the communication is embedded in order to produce and interpret the content that is being shared.

In face-to-face interaction, communication is multimodal, in the sense that it involves activation through more than one of the sensory modalities (hearing, vision, touch, smell and taste). When it comes to hearing and speech, both segmental and suprasegmental (prosodic) features of speech are involved. Gestures are involved both in touch and in visually shared information.

2.2. Modes of activation and representation

Besides multimodal communication, another important presupposition for our discussion, are the three basic semiotic means of information activation (representation) suggested by Charles Sanders Peirce; symbol, icon and index.

Index: Indexes involve activation or representation by making use of contiguity in space and time.

Icon: Icons involve activation (or representation) by making use of similarity.

Symbol: Symbols involve activation (or representation) by making use of conventional associations.

All these three modes of activation and representation are used in simultaneous and consecutive combination with each other, in both cognition and face-to-face communication. The symbols used are mostly vocal verbal expressions, while the icons and indexes are mostly gestures.

2.3. Theories of semantics for symbols, icons and indices

Theories of semantics have almost exclusively been concerned with written or vocal verbal symbols (words and combinations of words). Other types of symbols, icons and indexes have rarely been considered. Some of the most common semantic theories for written and vocal words (morphemes, phrases, sentence) are:

1. Truth conditional semantics (applies primarily to sentences)

2. Common meanings, in the form of necessary and sufficient conditions (primarily applies to words and morphemes)

3. Basic meanings, in the form of basic exemplars or prototypes (primarily applies to words and morphemes) [3]

4. Meaning Potentials, in the sense of the potentially activizable information connected with verbal symbols – here the point of departure is the collection of all of a word's uses in individual and collective memory [4], [1] For collective memory, see for example the linguistic division of labor discussed by Putnam in [5]).

The primary question we want to address in this paper is the question of whether meaning potentials can be extended from symbols to icons and indexes. Can we, drawing on memory and perception, in analogy with symbolic meaning potentials, also assume that there are iconic meaning potentials and indexical meaning potentials?

A Meaning Potential (in the sense we take it here, which is different from the sense it is used in for example, [6], [7],), is a structured collection of uses of a symbol (word), that is relevant both for understanding and production in communication.

Meaning potentials, thus, provide an analysis of linguistic meaning in line with the suggestion made by Wittgenstein [8] of seeing the meaning of a word as the set of uses of the word. But it is also in harmony with Vygotsky's suggestion that children learn language by learning linguistic labels (pseudo concepts) that are filled with content, as they successively learn to use these labels in different contexts [9]. The collection of uses forms the meaning potential of a word. Part of our linguistic competence is learning to activate (or actualize) this potential as triggered by different contextual features, such as the collocations (other words and morphemes), that a particular word is (often) combined with or the social activities in which the word are used.

The collection of uses (meaning potential) as stored in memory can then become a basis for a polysemy structure (analogous to what one might find in a dictionary) that is upheld by association with relevant contextual features like collocations and social activities. The meaning potential can also become the basis for cognitive processing, which can produce prototypes (typical examples), where they are relevant or necessary and sufficient conditions, where they are relevant, or both of these, when that is relevant. The cognitive processing is guided by cognitive operations supporting discrimination (analysis) and combination (synthesis), compression and abstraction of content, including such processes as contiguity abstraction and similarity abstraction and refinements of these. When the cognitive operations become associated with linguistic markers, we will refer to them as semantic-epistemic operations. Se also [4] and [1].

Let us now turn from the meaning potential of words and other symbols to a consideration of the role of meaning potentials for iconic and indexical gestures accompanying speech. As we shall see such gestures are often used to specify, highlight or illustrate features of the activated verbal (symbolic) meaning potential based content. Briefly, what happens is that in addition to the content activated by the words, the gestures activate additional content dependent on similarity (icons) and/or continguity relations (indices).

For example, iconic gestures might add illustrative pantomimes or metaphorical content and indexical gestures might add pointing to specific concrete or abstract locations or metonymic content.

In production, the gestures indicate, display or signal relevant information by making use of similarity and contiguity, often related to the verbal content. Similarly, in understanding, we interpret relevant information by similarity or contiguity, often in relation to the vocal verbal symbolic content.

2.4. Communication in context

Both of these processes – production and understanding – involve use of context as a resource for activation and contextual adaptation, accommodation, actualization and determination of content.

In fact, communication is always dependent on context for content, behavior action and type of interaction. The status and functions of new contributions are continuously being shaped by dimensions of context , such as:

- the physical environment
- the culture, the language, the current organizational setting
- the current social activity/activities
- the activity roles of the communicators
- the various traits of the communicators; gender, age and other psychological, social and biological properties
- the current contribution (compositionality)
- the currently preceding and/or simultaneous contributions (co-construction)
- other informative actions and behavior by the communicators
- the currently activated but also the potential shared background of the communicators (their "common ground")

The dimensions and features of context mentioned in this list form a background for pointing to two basic types of contextual determination of content

(i) Compositionality (combinability) in a wide sense

What we have in mind here is the contextual determination of the content of a multimodal contribution by drawing on the combined activation of several or all communicative features of the units (combining words and gestures) occurring in the same contribution. This is the issue we are discussing in this paper.

(ii) Co-construction

Here we move our contextual window from the content of a contribution of a single communicator to contextual determination of content, by drawing on the combined activation of several communicative contributions (mostly consecutive) from different communicators. This issue we will return to in future work.

2.5. Levels of awareness and intentionality

Our analysis also takes into account the fact that communication takes place on several simultaneous levels of awareness and intentionality. To facilitate analysis, we distinguish the following three levels on what basically is a continuous scale (cf. [10]).

Indicate (being informative)
Display (showing)
Signal (showing that you are showing)

These three levels can be combined with the three Peircean types of representation (index, icon, symbol) in the following manner, where all possibilities can occur but we have only indicated the most frequent cases in face-to–face communication.

Table 1. *Levels of awareness and intentionality and types of representation.*

	index	**icon**	**symbol**
indicate	Vocal segmental, Gesture, Prosody		
display	Vocal segmental, Gesture, Prosody	Vocal segmental, Gesture, Prosody	Vocal segmental, Gesture, Prosody
signal	Vocal segmental, Prosody, Gesture, Prosody		Vocal segmental, and prosodic verbal, Gestural verbal

The table shows how the three means of expressions (words, gesture and prosody) are typically related both to the three levels of awareness and intentionality and the 3 types of representation.

3. Meaning potentials, multi-representational and multimodal contributions

Using the background introduced above, we now want to discuss in what sense there can be meaning potentials not only for verbal symbols, but also for accompanying gestural indeces and gestural icons.

We want to do this by discussing what could be meant by these three types of meaning potentials, and then discussing what could be meant by combining them

3.1. The meaning potentials of symbols (words)

A meaning potential of a word can be organized into a polysemy compatible structure sustained by collocations related systematically to encyclopedic (including iconic and indexical) information. As an example, we present a sketch of the meaning potential of the word *tree* below

Tree: **Meaning potential: Polysemy + collocations:**
Source: http://oxforddictionaries.com/definition/english/tree [11]

noun

1a woody perennial plant, typically having a single stem or a trunk growing to a considerable height and bearing lateral branches at some distance from the ground. (in general use) any bush, shrub, or herbaceous plant with a tall erect stem, e.g. a banana plant.

2 a wooden structure or part of a structure.
archaic or literary the cross on which Christ was crucified.
archaic a gibbet.

3 a thing that has a branching structure resembling that of a tree.
(also tree diagram) a diagram with a structure of branching connecting lines, representing different processes and relationships.

verb (trees, treeing, treed) [with object]

1 North American force (a hunted animal) to take refuge in a tree.
informal, chiefly US force (someone) into a difficult situation.

2 as **adjective** treed (of an area) planted with trees sparsely treed grasslands

Collocations

- Decision tree
- Solution tree
- Tree diagram
- Elm tree
- Fruit tree
- Christmas tree

The meaning potential also includes and integrates encyclopedic meaning so no systematic distinction is made between lexical and encyclopedic meaning.

- Source Wikipedia – Encyclopedia [12]:
- In botany, a **tree** is a plant with an elongated stem, or trunk, supporting leaves or branches.
- In some usages, the definition of a tree may be narrower, including only woody plants, only plants that are usable as lumber, only plants above a specified height or only perennial species. At its broadest, trees include the taller palms, the tree ferns, bananas and bamboo.
- In its broadest sense, a tree is any plant with the general form of an elongated stem, or trunk, which supports the photosynthetic leaves or branches at some distance above the ground.[6][7] Trees are also typically defined by height,[8][9][10] with smaller plants being classified as shrubs,[11] however the minimum height which defines a tree varies widely, from 10 m to 0.5 m.[10] By these broadest definitions, large herbaceous plants such as papaya and bananas are trees, despite not being considered as trees under more rigorous definitions.[3][5][12][13][14][15]
- Another criterion often added to the definition of a tree is that it has a woody trunk.[10][16][17] Such a definition excludes herbaceous trees such as bananas and papayas. Monocots such as bamboo and palms may be considered trees under such a definition.[18] Despite being herbaceous[19][20] and not undergoing secondary growth and never producing wood,[21][22][22][23] palms and bamboo may produce "pseudo-wood" by lignifying cells produced through primary growth.
- Aside from structural definitions, trees are commonly defined by use. Trees may be defined as plants from which lumber can be produced.

Finally the meaning potential of a symbol can also include iconic and indexical information and contextual information, over and above that given by collocations.

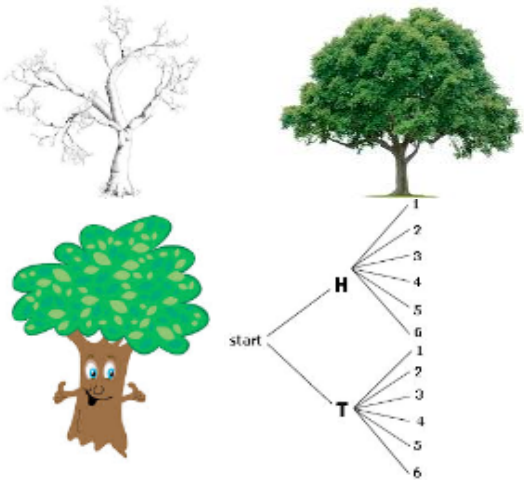


Figure 1: *Iconic information in meaning potential of tree.*

3.2. The meaning potentials of iconic gestures

The meaning potential of an icon relies on similarity, i.e. homomorphic-isomorphic relations that can be used both for production and understanding; activated by cognitive processes and semantic-epistemic operations triggering cognitive specification from memory or perception.

Let us first consider what might be the meaning potential of an icon without regard for context [13]. We explored this topic by asking a panel of judges to give interpretations of an iconic tree gesture, which consisted of: Both hands lifted in front of face, palms about 15 cm apart, turned towards each other, then hands coming apart and then together forming a circle, then both hands moving down in parallel.

Below are the interpretations of these iconic gestures as given by the panel of judges.

A. Showing a shape – possibly woman
B. A man or a person
C. Round a the top getting thinner – showing form
D. Tree
E. Showing the shape of something
F. Showing the form
G. Female earth mother (showing hip rouding)
H. "this shape"
I. quite
J. "symbolizing a woman/female body"
K. a tree
L. narrow it down

If we take these responses as indications of the meaning potential of the exhibited gesture, we can see that what seems to be going on is an activation of shapes from memory that are similar to the gesture.

As we can see, the meaning potential of a decontextualized iconic gesture, in general, seems more open and less structured than the meaning potential of a decontextualized verbal symbol. A circular movement of a hand or a finger can be similar to many things and we need either to add conventionalization or context, or both, to arrive at a more limited specific type of content. Deaf sign language has many examples of how iconic signs become conventionalized, that is, they combine iconic with symbolic representation and in this way can be used to activate more specific content.

For iconic signs that are less conventionalized than deaf sign language, context is needed to guide the users in what features, of the content being shared, are the relevant ones to focus on for the similarity based abstraction and activation of information. If the shared contextual information is not sufficient for this, there is a risk that the content activated will not be shared. This can clearly be seen in the variety of the responses presented above, where no shared context was provided.

In face-to-face multimodal communication, the most important context for iconic gestures is usually the content of the vocal verbal messages they co-occur with.

3.3. The meaning potentials of indexical gestures

To get an idea of the meaning potential of decontextualized indeces, let us consider two examples of indexical gestures: (i) a pointing index finger or (ii) a smile. As with iconic gestures, the meaning potential of these gestures, without conventionalization or context, will allow for a too large number of information activations.

What is being pointed to by an index finger is in no way easily restricted, since it could be pointing to both concrete and abstract entities. What is being expressed by a smile is more restricted, but in itself allows for many interpretations, like friendliness, shame, fear, ingratiation, happiness, contentment, malevolence etc.

As with iconic gestures, context is needed to determine what the contiguity relation activated by the gesture should apply to.

3.4. Multimodal combinations of symbols, icons and indices in face-to-face communication

(i) Symbol with symbol

The first combination to consider is perhaps the multimodal combination of a vocal verbal symbol with a gestural verbal symbol. Such combinations are common in giving communicative feedback in English, where, for example, the vocal verbal symbol *yeah* is often accompanied by a gestural verbal symbol, affirmative *head nod,* providing a multimodal combination of a vocal and a gestural symbol, both expressing assent and affirmation, the function of which is a reinforcement of the affirmation. The same would happen, in English, if the vocal verbal *no* is combined with a gestural verbal *head shake.*

(ii) Symbol with icon

Let us now see what happens if the meaning potential of an iconic gesture is combined with the meaning potential of a

word. Let us consider an example from a discussion about Nature.

Example 1:

B: he was there // with his senses and open to it just then // maybe sitting on his tractor //and
D: yes
B: he probably didn't then // but normally [it is probably (…)]
C: [(…) forerunners] with with modern tractors // with // air condition // radio // and headphones // machine panel
D: but // but surely // e // e // surely // there is someone who has a //quick // association // e ö with a // **fruit tree (+ iconic gesture)** blossoming // and who sits driving a tractor // and turns around

Note: (// = pause, [ ] = overlap, (…) = inaudible speech)

In this example, the iconic gesture triggers a cognitive similarity specification, operating on the actualized content of the word *tree* and other perceptual memories related to this word.
The gesture highlights the shape of a blossoming fruit tree and in doing this also emphasizes and specifies the tree and the shape of the tree.

(iii) Symbol with index

As with iconic gestures, the context of an indexical gesture will often be given by simultaneously produced vocal verbal content. For example

1. The house is over there, accompanied by *pointing* gesture

The gesture specifies direction to the location of the house by contiguity and the verbal element tells us what is to be located.

2. *I am happy to see you*, accompanied by a *smile*

Here the smile indicates an inner state of happiness expressed by the word *happy*.

In both cases, the gestures (the pointing finger and the smile), that could potentially have many other meanings, trigger an epistemic contiguity operation which further specifies the content of the vocal verbal symbol.

(iv) Symbol with both index and icon

Often, vocal verbal symbols are combined with both iconic and indexical gestures, which can occur either separately or as simultaneous features of one gesture. Let us consider some examples.

Example 2.

A: *I have tried to start to study also English, so that I wont* ***forget*** **(the word *forget* is combined with an indexical/iconic gesture of a circling index finger pointing to the temple of the head)**

The activated meaning potential of the word *forget* here provides the contextual content basis for the gesture, which in itself combines indexical and iconic features.

The indexical features (contiguity in space and time) of the gesture locates "the forgetting" in the head. Here perception of this gesture and memory interact in giving further associations to cognitive processes. The iconic features of the gesture, "a circling motion", simultaneously with the indication of the location, highlights a memory problem (circling – not finding).

To some extent, this particular association between the gesture used (the circling finger) and a cognitive problem has been conventionalized, which can be seen when we asked a panel of judges to give interpretations of circling index finger pointing to head and the majority all indicate a cognitive problem of some sort.[13].

Description of "forget" gesture:
Preparation: lifts hand up towards head
Stroke: index finger points to head, circular movement
Retraction: hand goes back to lap

Suggested meanings by panel of judges:

A. crazy (in the head) or confused /about self
B. "I am confused"
C. Don't understand – crazy/nuts
D. ?
E. I'm crazy /confused
F. You have a hole in your head, you know = are stupid /don't understand
G. I am confused!
(pointing to ear and circling to show confusion)
H. "this person is crazy"
I. thinking all the time
J. mind-bogging
K. "my head is going round" = "cocco
L. cannot remember, or cannot think sth up
M. Hopeless to talk to

If we consider the contextual influence on the meaning activation of the multimodal contribution, we see that we have in this case is a combination of

(i) activation of the symbolic meaning potential of the word *forget*, which is contextually determined by the fact that it occurs in the activity context of a discussion on learning and is a collocation of *not forget*.

(ii) a gesture providing an indexical highlighting of the locus of forgetting and an iconic highlighting of a dynamic circle, which can display some type of cognitive problem.

The combined multimodal effect of the gesture will be to highlight and emphasize the locus of forgetting in the head.

4. Summary and concluding discussion

We have seen that multimodal face-to-face communication not only involves a combination of information in many modalities, but also a combination of several modes of representation on several level of awareness and intentionality.

What we frequently have is a combination of vocal verbal symbols with gestural icons and indices. However, vocal verbal symbols can also be combined with gestural verbal symbols, as is the case in communicative feedback, where, in English, words like *yes* and *no* are combined with head nods and head shakes. The most common effect of a combination of a vocal verbal symbol with an iconic or indexical gesture is that features of the activated symbolic content are specified, highlighted or illustrated by making use of cognitive semantic-

epistemic operations relying on similarity (homomorphism) and/or contiguity relations. When vocal verbal symbols are combined with gestural verbal symbols, the effect is rather one of reinforcement and emphasis.

The framework we have proposed, thus, provides some steps towards a cognitive, holistic semiotic theory of the semantics/pragmatics of multimodal contributions to interactive communication. We have suggested that communication should be seen as involving productive activation and receptive coactivation of shared content (information/understanding), drawing on contextual resources determining the meaning potentials of symbols (words), but also of icons and indices, making use of cognitive operations combining activation of conventional information with activation making use of similarity and contiguity relations, helping to determine the symbolic content.

We have also seen that the meaning potentials of symbols, icons and indices vary in how specific and structured the information is that they enable activation of. Conventionalization, in general, seems to make the activizable information more specific and structured, so that decontextualized symbols have more fine grained meaning potentials than decontextualized icons and indices. It seems likely that the same difference might also hold true for icons versus indices. The homomorphism of icons puts more restrictions on what information can be activated than the contiguity of indices.

In all cases, activation of meaning potentials requires activation of contextual resources to acquire a more determinate actualized meaning. Since the meaning potentials of icons and indices are more open ended than the meaning potentials of symbols, this need is stronger for icons and indices than for symbols. In this way, iconic and indexical gestures have a natural functional role to play as coverbal reinforcers and specifiers of features of content in activated symbolic verbal meaning potentials.

Thus, our analysis provides a basis for a rethinking not only of the "compositionality" of symbols (symbol + symbol) in terms of a combination of symbolic meaning potentials under contextual enablements and restrictions, but also for considering the combination of symbols (words) with icon and index (usually gestures) under contextual enablements and restrictions.

Finally, we have noted that meaning potentials with the aid of cognitive (semantic-epistemic) operations can be used not only as a basis for meaning determination and meaning actualization in context, but also to operate innovatively on shared information, creating new generalizations, prototypes and metaphors; sometimes reinforced by innovative gestures.

5. References

.

[1] J. Allwood, "Meaning Potential and Context: Some Consequences for the Analysis of Variation in Meaning." In H. Cuyckens, R. Dirven and J. R. Taylor, (eds.) *Cognitive Approaches to Lexical Semantics*. *Mouton de Gruyter*. pp. 29-65, 2003.

[2] C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, vols. 1–6, S. Hartshorne and Paul Weiss, eds., vols. 7–8, 1958, A. W. Bukrs (ed), Harvard University Press, Cambridge, Massachusetts, 1931-35, 1958.

[3] E. Rosch, "Cognitive representations of Semantic Categories". *Journal of Experimental Psychology:* General 104 (3): 192–233, 1975.

[4] J.. Allwood, "Semantics as Meaning Determination with Semantic Epistemic Operations." In J. Allwood and P. Gärdernfors (eds.) *Cognitive Semantics*. Amsterdam: Benjamins. pp. 1-18, 1999.

[5] H. Putnam, "The meaning of 'meaning." In *Philosophical Papers*, Vol. 2. *Mind, Language and Reality*. Cambridge University Press, 1975/1985.

[6] M. A. K. Halliday, *Explorations in the Functions of Language* . London : Edward Arnold , 1973.

[7] M. A. K. Halliday,. *Language as Social Semiotic: The social interpretation of language and meaning*. London : Edward Arnold, 1978.

[8] L. Wittgenstein,, *Philosophical Investigations*. Oxford: Blackwell Publishing, 1953.

[9] L. Vygotsky, *Thought and Language*. In A. Kozulin (Ed. and Transl.). Cambridge, MA.: MIT Press, 1986.

[10] J. Allwood, "A multidimensional activity based approach to communication." In I. Wachsmuth,, J. P. de Ruiter, P. Jaecks and S. Kopp (eds) *Alignment in Communication.* Amsterdam: John Benjamins, pp. 33-55, 2013.

[11] http://oxforddictionaries.com/definition/english/tree

[12] Wikipedia – Encyclopedia. https://en.wikipedia.org/wiki/Tree

[13] E. Ahlsén, and J. Allwood, "What's in a gesture?", *Allwood, J,. Ahlsén, E., Paggio, P. (2013). Proceedings of the Fourth Nordic Symposium on Multimodal Communication, Nov 15-16, University of Gothenburg. NEALT Proceedings Series No. 21.*. Linköping: Linköping University Electronic Press, pp. 13-20, 2013.

A User Study on Awareness Signals for Social Communication

Dimitra Anastasiou[1], *Wilko Heuten*[2], *Susanne Boll*

[1]Luxembourg Institute of Science and Technology
[2]OFFIS – Institute for Information Technology, Escherweg 2
[3]Media Informatics and Multimedia Systems
University of Oldenburg, 26121 Oldenburg, Germany
dimitra.anastasiou@list.lu, wilko.heuten@offis.de, susanne.boll@informatik.uni-oldenburg.de

Abstract

Very often today people are depending on distant social communication to maintain contact with their working groups or families. This distant communication happens often very abruptly without any awareness signal in advance. This paper presents a Wizard-of-Oz user study based on awareness signals, specifically illumination and sound effects, which were triggered by an experimenter before the communication started. Participants had to test a distant vs. a close light vs. a sound effect vs. combination of light-sound vs. absence of any signals. Although the distant light was in the periphery of the focused attention of the users, it was generally better accepted, though less perceptible than the close light. Promising results towards "peripheral awareness" show that the existence of triggered awareness signals in the unobtrusive periphery transmit a communication message fluently to the users.

Index Terms: awareness, human-computer interaction, interaction design, social communication.

1. Introduction

Nowadays synchronous (e.g., telephone, video-conference systems) and asynchronous (e.g., email, social networks) means help us communicate with our distantly separated peers. However, this distant communication is not as productive and effective as face-to-face communication and it often happens abruptly and obtrusively. At workplace when teams are co-located, spontaneous communication occurs very often at a daily basis: people meet at the coffee corner or have lunch together. However, nowadays due to globalization, the teams are often distributed over branch offices located in different cities and countries.

Let us imagine that at the foyer of a company there is an ambient display and we are currently passing by. In a branch office in another city a colleague does the same. It would have been nice if both colleagues would be aware of each other and have some social communication. Before a video communication software pops up, what kind of signals would users expect as an output from the display? It could be, for instance, a pulsing background light, a soundscape with increasing volume as we come closer or even an avatar, which welcomes us and introduces to the display's functionalities. As an input, there could be an in-air hand gesture, voice, but also raw sensor data, such as spatial distance, etc. Some of these modalities are more implicit, some more explicit; the transition from implicit to explicit communication should be transparent, but fluent. In our opinion, awareness signals before the beginning of communication would make the transition from the actual activity state to the communication state more fluent and in addition, would preserve privacy.

Our research is on facilitating spontaneous and informal communication in spatially distributed working groups by exploiting smart environments and ambient intelligence. In the project SOCIAL (Van de Ven et al. [1]) we focus on this research goal through the following three key steps:

1. Detection of situations with the potential for spontaneous informal communication;
2. Representation of these situations appropriately to distant users;
3. Enabling them to engage in communication spanning multiple spatial locations.

The above steps include the perception of the current potential communication situations, the transparent and privacy-preserving detection of instances of situations, representation of formalized behavioural cues in distributed setting, and last but not least, human-computer interaction (HCI) methods. The detection step includes representation and reasoning about the situational context. This requires a formal language to describe specific situations of interest, available knowledge, e.g., abstracted perceptions of situation context, and the behavior of the system. [1] applied methods from the field of qualitative spatio-temporal representation and reasoning (QSTR).

In this paper we focus particularly on the HCI methods for social communication and their requirements, such as implicit/ implied communication, intuitiveness, and unobtrusiveness. A pilot study is a first step towards exploring which signals are more appropriate for designing an awareness-communication system which fills these requirements. The paper is structured as follows: in Section 2 we present some related work of this interdisciplinary field and in Section 3 we discuss the user study, including the set-up (3.1), hypotheses and experimental methods (3.2) as well as its results (3.3). We have a short summary and discussion in Section 4 concluding the paper with a few future prospects in Section 5.

2. Related Work

As our research is interdisciplinary covering, among others, Sociological Studies, Ambient Intelligence, and Awareness Systems, here we present only a few related work of these fields to set the scene where the article belongs to. Kiesler and

Cummings [2] reviewed the term of *proximity* in work groups since the 60s and concluded that for distributed work groups, the use of communication technology is likely to be most successful when work groups are cohesive, i.e. they have already forged close relationships, so that the existing feelings of alliance or commitment sustain motivation. More recently and in the domain of Internet of Things (IoT), Atzori et al. [3] claim that in analogy with the human evolution from *homo sapiens* to *homo agens*, we may talk of an evolution path from a *res sapiens* (smart object) to a *res agens* (an acting object) and even to a *res socialis* (social object). The *res sapiens* communicates with the external world by relying on web protocols and communication paradigms by the current Internet of Services, while the *res socialis* refers to an object that is part of and acts in a social community of objects and devices.

Awareness systems, as a subfield of Ambient Intelligence, can be broadly defined as "systems intended to help people construct and maintain awareness of each other's activities, context or status, even when the participants are not co-located" (Markopoulos et al. [4]: v). There have been many systems in the past, the so-called *media spaces* connecting separate places, such as *Portholes* (Dourish & Bly [5]) and *Telemurals* (Karahalios et al. [6]). Moreover, the *Hello.Wall* and *Personal Aura* artefacts by Streitz et al. [7] emitted awareness information between distributed team members. *Hello.Wall* was an ambient display that emitted awareness information via different light patterns. Personal Aura (PA) enabled persons to indicate their "professional role" and "availability" to remote team members. PA consists of a reader module and an ID stick containing a unique identity and optional personal information.

Our research work, similarly as for [5], [6] and [7], can be categorized under "workspace awareness" systems. Gutwin & Greenberg [8] defined workspace awareness as "the collection of up-to-the-moment knowledge a person uses to capture another's interaction with the workspace". In 2001 Gutwin & Greenberg [9] created a workspace awareness framework with three aspects of: i) component elements (answer Wh questions), ii) mechanisms to maintain it (gather perceptual information), and iii) its uses in collaboration.

As far as auditory awareness signals are concerned, work goes back to middles 80s, when Sumikawa [10] provided guidelines for the integration of audio cues into computer user interfaces. In late 90s the Audio Aura system (Mynatt et al. [11]) provided serendipitous information tied to physical locations and delivered via portable wireless headphones. The PANDAA system (Sun et al. [12]) was a zero-configuration spatial localization system for networked devices based on ambient sound sensing. Ambient sounds, such as human speech, music, foot-steps, finger snaps, hand claps, or coughs and sneezes, were used to autonomously resolve the spatial relative arrangement of devices in a ubiquitous home environments using trigonometric bounds and successive approximation.

More recently Kainulainen et al. [13] presented guidelines regarding six common auditory techniques: speech, auditory icons, earcons, music, soundscapes, and sonifications and designed a general structure of an audio awareness architecture following the agent-evaluator-manger principle.

Within the project SOCIAL, Sartison [14] conducted an online survey and interviews with 23 participants to set the requirements for designing a stationary prototype which exchanges unobtrusive audio messages with the users as awareness signals. Participants had to evaluate a speech message vs. an auditory icon (sound of a coffee machine) vs. an earcon vs. a soundscape (cafeteria environment). The speech message was ranked as the most informative, but also the most obtrusive one. The auditory icon occupied the second place with regards to its perception, following very closely the speech message. Based on these user requirements, [14] developed a stationary prototype which automatically sends audio messages to users (mobile) based on their spatial location and their calendar availability. She also developed an mobile application to set up custom settings and the assignment of audio signals to a specific person.

As for visual signals, we particularly focus on illumination. Müller et al. [15] presented six examples of ambient light information displays, which address humans' perception abilities to gain cues from the periphery instead of attracting the user's visual focus. Our future system distinguishes from [15] as it will not be an information display system, but it will explore peripheral awareness through visual cues. Ehrhardt [16] designed a social communication vase with bubbling and colour-changing water based on the status of the social communication between remote people: orange colour when the situation for communication is detected, green colour to give the consent for communication, and red to decline it. The prototypes [14] and [16] used Raspberry Pi and Arduino respectively.

For the evaluation of our awareness system, we considered some of the heuristics of Mankoff et al. [17]: i) peripherality of display, ii) match between design of ambient display and environments, iii) easy transition to more in-depth information, iv) visibility of state, and vi) aesthetic and pleasing design.

3. User Study

In this section we discuss the study's set up (3.1), our hypotheses along with the post-study questionnaire (3.2), as well as the most significant results (3.3).

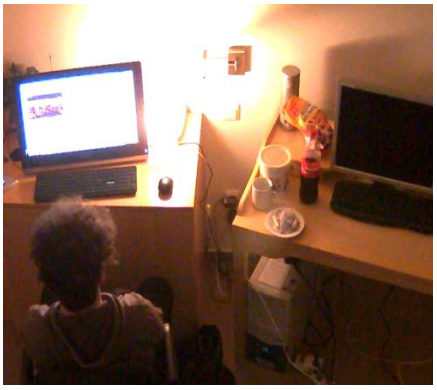
3.1. Set up

The user study took place in April-May 2015 at a lab at the University of Oldenburg. The goal of the study is to get initial results about the perception and overall acceptance of close/peripheral as well as visual/auditory awareness signals. 17 subjects (11 female, 6 male, mean age=25) participated in the study. Apart from one participant who has never used video communication software before, most of them were computer-savvy (but not computer science students). Each experiment lasted about 45 minutes and had two parts: i) a WoZ experiment and ii) filling in a user experience post-study questionnaire. As for the former part, the participants were asked to sit at a desk and watch a music video on a PC monitor at low volume; they were offered to have coffee and sweets during the video watching. With this setting we aimed at simulating a working environment, though not tied with a hardly concentrating job task, but rather a coffee break. They were informed that various signals, like light and sound, would appear in the room, without the experimenter pointing or verbally explaining the signal's exact output source. Should the participant notice a signal, (s)he should wait 5 secs and then call the experimenter on *Skype*; the addresser-addressee process and the communication mean *per se* was not the focus in this study. The experimenter was at a surveillance room and

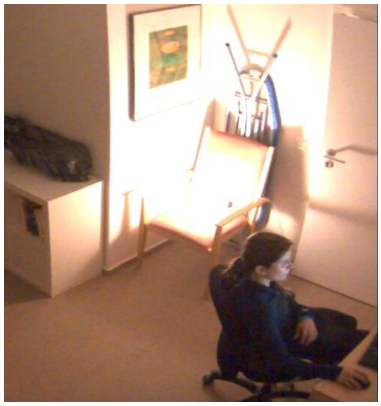
triggered the light and sound remotely. There were five experimental conditions tested:

i) Close light (lamp was next to the PC);
ii) Distant light (lamp was on a chair on the participant's left side);
iii) Sound (output of a wall-mounted speaker);
iv) Combination of sound and light;
v) Absence of signals.

The lamp used in the first, second, and fourth condition is a small lamp, ca. 30 cm high and it was illuminated white; there was not any pulsing light or other light patterns used. The investigator turned the lamp on and off with a time interval of 5 secs. The distant lamp was about 90º at the left side of the participant. This means that the actual signal perceived by the participant was a change in ambient illumination and their own shadow cast over the work area. The close light was in direct line of sight of the participant. For these reasons, we consider the distant light as a peripheral signal and the second as not peripheral signal. The lamp selected for the fourth condition was also random; for half of the participants it was the close lamp triggered and for the other half the distant one. The sound was output from a wall-mounted speaker, also about 90º at the right side of the participant, thus a peripheral audio signal. Pictures 1 and 2 show the setting of the experiment and precisely the close (Pic. 1) vs. distant lamp (Pic. 2).



Picture 1. Close light as awareness signal



Picture 2. Distant light as awareness signal

3.2. Hypotheses

In usability testing, questionnaires or subjective evaluations are used to learn from the users what a usable system is. To system designers, subjective evaluations may provide more informative data of system functionality than objective performance measures, since they focus on the user's first-hand experience. In our study we selected two subjective evaluation measures: a closed-ended questionnaire and a think-aloud protocol at the end of each experiment. In this sub-section we present our hypotheses with regards to spatial position of the awareness signal, overall user acceptance as well as the transition from awareness to communication. After each hypothesis, the questions of the questionnaire selected to test those hypotheses are presented.

Hypothesis 1 (Spatial position): *The spatial position of the signal's source influences its perception. A signal close to the user is easier and faster receipted than a distant one.*

The questions in the questionnaire that test this hypothesis are:

Q1	Does the spatial position of the lamp influence its perception?
Q2	How easily perceptible was the signal?

Hypothesis 2 (User acceptance): *The signal source that is close to the communication medium is better accepted by the users than the distant one.*

Sub-hypothesis 2a (Effectiveness): *The combination of two or more awareness signals is more effective than a single signal.*

The questions to test hypothesis 2 and 2a are:

Q3	Which of the following signals do you prefer?
Q4	How did you like the design/form of the signal?
Q5	Evaluate the idea of using light, sound, and the combination of light and sound as awareness signals.

Hypothesis 3 (Transition): *Peripheral signals provide a more fluent transition to communication than close signals.* The relevant questions to test this hypothesis are:

Q6	How gradual was the transition from the task to communication?
Q7	How much did the signal distract you from your task?

In statistical terms, we have three *discrete dependent variable* and four *discrete independent variables*; each independent variable has five *levels* on Likert scale (1-5)/ordinal variables. For Hyp.1, we tested 4 independent variables (absence of signals was excluded). We had a *within-subject* design, i.e. each user performs under each different condition. In order the design not to suffer from transfer of learning effects, we randomized the order of the conditions for each participant.

Discrete dependent variables	Perception, user acceptance, transition
Discrete independent variables	Close vs. distant light vs. sound vs. combination of sound and light vs. absence of signals.

3.3. Results

The results of our study are presented in the order of the hypotheses presented above.

Hypothesis 1 (Spatial position): In a dichotomous yes/no question (Q1), 88,24% of the participants stated that the spatial position of the lamp influences its perception. A significant percentage of 11,76% did not share this opinion, showing that awareness signals in the periphery does not seem to affect its perception negatively based on the user's experience. Diagram 1A presents the options along the Likert scale.
As far as the perception of signals is concerned (Q2), the close light was evaluated as the most easily perceptible signal with 94,12% (scale 5-strongest perception) followed by the distant light with 64,71% (scale 4). Comparing the close light with the sound in particular, on one hand, the close light raised the strongest awareness of most participants (MD=5, σ=0,24, Var=0,06). The sound, on the other hand, showed a much higher standard deviation and variance (MD=4, σ=0,9, Var=0,81). Based on the think-aloud protocol, a participant said that he perceived the light much faster than the sound, while another one mentioned that the sound has to be repeated to be more perceptible. Diagram 1B depicts in a boxplot the min, max, Q1, Q3 and MD values of the five signals.

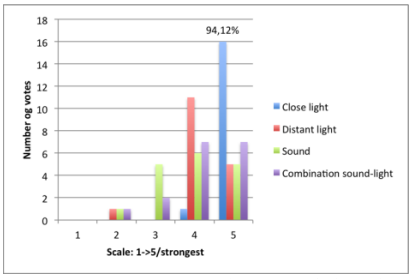


Diagram 1A. Perception of various signals based on their spatial position

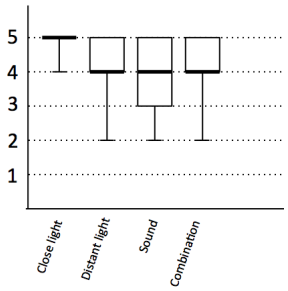


Diagram 1B. Boxplot about the perception

Hypothesis 2 (User acceptance) and sub-hypothesis 2a (Effectiveness): As for the overall preference of signals (Q3), Diagram 2A shows that most participants (64,71%) selected the situation-dependent option and not the combination of the signals (29,41%)[1]. The former means that either light or sound is selected as a signal based on a specific situation, e.g. in a loud environment, light is more appropriate, whereas in a very bright environment, sound is rather appropriate. Moreover, the situation-dependent option makes the system accessible for people with disabilities; a participant was a Sign Language Interpreter and mentioned that for the deaf, light signals lit up on every door when there is a knock on the entry door. Remarkably, light was ranked second with a big gap from sound (difference of 52,94%). º
With regards to aesthetic design of the signals (Q4), the distant light was ranked higher (MD=5), whereas the sound was less accepted (MD=2). This might be due to the kind of the selected sound effect (whispering "Psst..psst" sound); from the think-aloud protocol we deduce that participants would rather prefer a sound similar to an alert tone, a bell sound, or typical mobile phone tones that most people are familiar with. As the user acceptance is subjective and very much dependent on the selection of the study's triggered signals, we also report some statements from the participants (Table 1).

Table 1. *Statements from the think-aloud protocol about light, sound, and the combination*

The light is more user-friendly and discreet than the sound; you can easily blend it out in order to watch the video. The sound is always the same. You get frightened by the close light.
The close light was too penetrative.
If the close light was brighter, I would prefer that.
If you are concentrated, you don't perceive the distant light strongly.
The sound hacks me off, as it should happen often in order to be perceived.
You can mistake the awareness sound with another sound.
One signal is actually sufficient, as the combination leads to stimulus satiation.

As far as the evaluation of the idea of using light, sound, and the combination as awareness signals is concerned (Q5), participants were asked to compare their familiarity, interest and necessity (questionnaire's pre-defined answers). As expected, the sound was most familiar due to auditory signals known from mobile phones. The awareness with light was ranked equally interesting and necessary with the combination.

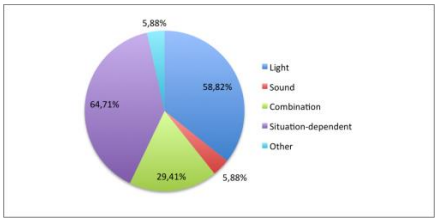


Diagram 2A. General preference of signals

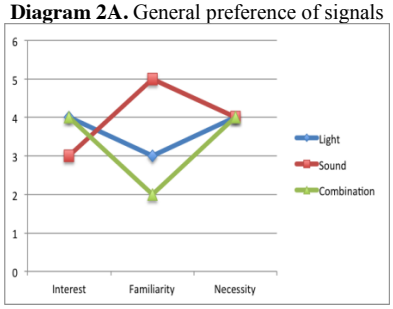


Diagram 2B. Evaluation of the idea of using awareness signals

[1] The sum of the votes is over 100%, as this question allowed multiple answers.

Hyp. 3 (Transition): Diagram 3A shows the transition with the absence of signals being ranked as the most abrupt (47,06%-scale 1), whereas through the distant light as the most fluent (52,94%-scale 5) transition. Regarding the results of the distraction from the actual activity, the distant light seemed to distract less (Var=1,28, σ=1,13) than the close light (Var=2,62, σ=1,62). The fact that the distant light distracted less justifies the fact that the distant light seemed to provide a more fluent transition to communication, as evaluated by the users. As the Diagram 3B boxplot shows about the transition from awareness to communication through the different options, the MD was the same for close and distant light and sound (MD=4), while it was lower for the combination (MD=3) and very low for the absence of signals (MD=1).

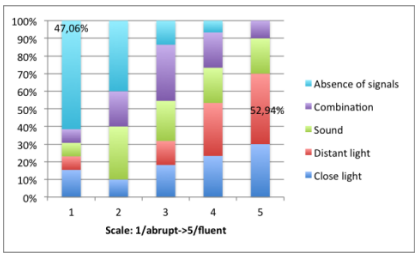


Diagram 3A. Transition from awareness to communication

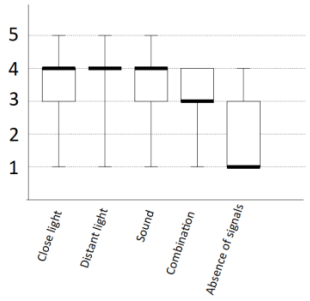


Diagram 3B. Boxplot about the transition

In addition to the results based on the hypotheses, we did a video annotation in order to test the viewing position of the participants when the peripheral light was triggered.
We deduce that 76,47% of the participants did not turn their head (focus) to look at the peripheral light. The remaining 23,53% looked at the peripheral light and did that even repeatedly after they noticed the signal. One out of 17 participants did not notice the distant lamp *per se*, although he realised that there was an ambient light. Moreover, one of the participants who looked at the distant lamp did that only after he called the investigator on *Skype*; that shows that awareness was raised before. Last but not least, only one of 17 participants looked at the wall-mounted speaker, when the sound was triggered.

4. Discussion

Non-verbal or implicit communication is very important in our frequent communication with our spatially distributed co-workers. This kind of communication includes the perception of the currently performed activity, behaviour or the presence of other people.

In this paper we presented our pilot user study regarding awareness signals for social communication. We evaluated their perception, user acceptance, and transition to communication. The system to be designed is a workspace awareness system in which visual and/or auditory signals will notify the co-workers in spatially distant settings that at that moment there is an opportunity for communication. The results showed that in general, light was higher accepted than sound and definitely better than the combination of two signals. Although the close signal was more perceptible than the distant one, the peripheral signal was more highly accepted by the users regarding aesthetics, unobtrusivess, and provision of a fluent transition to communication. These results along with the fact that the majority of the participants did not turn their viewing direction to the distant light show that raising peripheral awareness is not only feasible, but also effective, privacy preserving, and more closely tied with implicit communication which is a crucial requirement for our system.

5. Future Prospects

Awareness systems should be able to capture the presence of other people or their performed activity. The future system should be unobtrusive, scalable, and customizable to the user's needs. For these and other reasons, the future interface should be multimodal in order to give the user the opportunity to intuitively choose the interaction mode and easily use this mode. So far in practice, there is unfortunately no technical support system for implicit communication between spatially separated people.
In the future, we would like to explore further possibilities of multimodal signals for awareness and communication systems. This is possible by interpreting social signals through the recognition of behavioral cues (Vinciarelli et al. [18]), such as facial expressions, head movement, body gestures, voice detection, and speech recognition. Last but not least, as far as the representation design of our future system is concerned and with the actual developments of the Internet-of-Things (IoT), the ambient display might be replaced with any smart object that is available in a pervasive (working) environment: coffee mug, desk, chair, whiteboard, flower pot, etc. For instance, Wallbaum et al. [19] developed an artificial social plant which enables users to keep track of a loved person throughout the day by unobtrusively visualizing the partner's current state of mind via different colors of the blossom.

6. Acknowledgements

We acknowledge German Research Foundation (DFG) funding for project SOCIAL (FR 806/15-1 | BO 1645/12-1). The first author was working on the project SOCIAL at the department of Media Informatics and Multimedia Systems, University of Oldenburg during the conduction of this study.

7. References

[1] J. Van de Ven, D. Anastasiou, F., Dylla, C. Freksa and S. Boll, "The SOCIAL Project. Approaching Spontaneous Communication in Distributed Work Groups" *Proceedings of Ambient Intelligence Conference*, 2015.

[2] S. Kiesler and J.N. Cummings, "What do we know about proximity and distance in work groups? A legacy of research", *Distributed work*, 1, pp. 76–109, 2002.

[3] L. Atzori, An. Iera and G.Morabito, "From "Smart Objects" to "Social Objects": The Next Evolutionary Step of the Internet of Things", *IEEE Communications*, 52, pp. 97–105, 2014.

[4] P. Markopoulos, B. De Ruyter and W. Mackay, *Awareness systems: advances in theory, methodology, and design*. Human-Computer Interaction Series. Springer London, London, 2009.

[5] P. Dourish and S.A. Bly "Portholes: Supporting awareness in a distributed work group" *Proceedings of CHI*, pp. 541–547, 1992.

[6] K. Karahalios and J.S. Donath, "Telemurals: linking remote spaces with social catalysts" *Proceedings of CHI*, pp. 614–622, 2003.

[7] N., Streitz, C. Röcker, T. Prante, R. Stenzel and D. van Alphen, "Situated Interarction with Ambient Information: Faciliating Awareness and Communication in Ubiquitous Work Environments" *Tenth International Conference on Human-Computer Interaction* (HCI International 2003). Citeseer, 2003.

[8] C.Gutwin and S. Greenberg, "Workspace Awareness for Groupware" *Proceedings of the Human Computer Factors in Computing Systems* (CHI 1996), ACM Press, 208–209. 1996.

[9] C.Gutwin and S. Greenberg, "A Descriptive Framework of Workspace Awareness for Real-Time Groupware". Computer Supported Cooperative Work, Kluwer Academic Press, 2001.

[10] D.A. Sumikawa, *Guidelines for the integration of audio cues into computer user interfaces*. Lawrence Livermore National Lab., CA, 1985.

[11] ED. Mynatt, M. Back, R. Want, M. Bear, and Ellis, JB, "Designing Audio Aura", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1998.

[12] Z. Sun et al. "PANDAA: physical arrangement detection of networked devices through ambient-sound awareness" *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 425–434, 2011.

[13] A.Kainulainen, M. Turunen, and J. Hakulinen, "Awareness information with speech and sound" *Awareness Systems-Advances in Theory, Methodology and Design*, pp. 231–256, 2009.

[14] M. Sartison, *Kommunikation über Audiosignale zwischen räumlich entfernten Gruppen* (*Communication by means of audio signals between spatially remote groups*), MA thesis, University of Oldenburg, 2016.

[15] H. Müller, et al., "Ambix: Designing ambient light information displays" *Proceedings of Designing Interactive Lighting workshop at DIS*, ACM, 2012.

[16] B. Ehrhardt, *Lenkung der Aufmerksamkeit auf eine potentielle Kommunikationssituation im Büroalltag (Awareness control to a potential communication situation at workplace)*, BA thesis, University of Oldenburg, 2015.

[17] J. Mankoff, A.K. Dey, G. Hsieh, J.A. Kientz, S. Lederer and M. Ames, "Heuristic evaluation of ambient displays" *Proceedings of CHI*, 169–176, 2003.

[18] A. Vinciarelli, M. Pantic, M. and H. Bourlard, "Social signal processing: Survey of an emerging domain" *Image and Vision Computing*, 27, 12, pp. 1743–1759, 2009.

[19] T. Wallbaum, J. Timmermann, W. Heuten and S. Boll, "Forget Me Not: Connecting Palliative Patients and Their Loved Ones" *Proceedings of CHI Extended Abstracts*, pp. 1403–1408, 2015.

Multimodal Representation of Shared Cultural Knowledge in a Creolophone Community

Melanie Anna Brück [1]

[1] University of Cologne, Germany
brueckm@uni-koeln.de

Abstract

This paper discusses how cultural aspects interact with both vocal and gestural features in multimodal communication, and, more specifically, how shared cultural knowledge influences the form and function of gestures. For this, the multimodal representation of shared cultural knowledge in Kreol Seselwa (KS), a French-based creole language spoken on the Seychelles, will be analysed. The domain of shared cultural knowledge investigated here is spatial orientation and conceptualisation, focusing on the three spatial Frames of Reference (FoR) defined as intrinsic, relative and absolute. Both elicited and semi-spontaneous data collected on the Seychelles show that one striking feature of this creolophone community seems to be a dynamic use of several FoRs in everyday communication. It will be illustrated that in KS it is the availability of culturally shared knowledge, amongst other factors such as modality and context, which influences the choice for a certain FoR. On the gestural level, the data show how culturally shared knowledge of Kreol Seselwa speakers is represented by phonological features as well as the use of abstraction in pointing gestures referring to existing places. Furthermore, the data illustrate the dynamics of merging deictic and iconic elements in gestures accompanying locally-anchored narrations and how this reveals aspects of shared background knowledge. The representation of shared cultural knowledge in KS across modalities emphasises the importance of interpreting multimodal data in the light of the micro-ecology of communication, taking both linguistic and extra-linguistic factors into account.

Index Terms: Spatial Reference, Frames of Reference, Multimodality, Shared Cultural Knowledge, Micro-ecology of Communication, Creole Languages

1. Introduction

1.1. Speech, Gesture, and Culture

Efron, [1] was one of the first researchers to systematically compare gesture use across cultures – a line of research being followed by several studies since then [2] – [8]. While often these comparisons have focused on the interaction between gesture and speech, some studies have also taken into account the interface of gesture and culture, thus considering not only speech but also "[g]estural practices as cultural tradition" [3, p. 328]. An important notion in the interface of communication and culture is the so-called micro-ecology of communication, i.e. the environment in which a communicative act is being performed. This includes all extra-linguistic factors that may influence the shape of language and gesture use in a certain community [3, pp. 305f.]. It is commonly acknowledged that human communication interacts with e.g. cultural, historical, social, political and ecological factors. The claim made in this paper is that shared cultural knowledge is one of these factors playing a role in the micro-ecology of communication, shaping not only speech but also gesture. Shared cultural knowledge can be seen as practices of knowledge organisation that are socially distributed and both created and interpreted by a certain community [9] – [12]. In other words, we are looking at the kinds of resources a society uses to link and anchor entities and concepts to one another. Such shared cultural knowledge and its representation in the visual modality has been described for both gesture systems [7], [13], [14] and sign languages [15] – [17]. Typical domains of shared cultural knowledge are for example kinship systems and person reference [18], [19], environmental knowledge [20] and medical knowledge [21], [22]. The domain of shared cultural knowledge investigated in this paper is spatial orientation and conceptualisation. The data analysis suggests that KS speakers apply a mix of strategies to refer to spatial setups and that it is the interaction with shared cultural knowledge which shapes the dynamics of speech and gesture interaction.

1.2. Spatial Reference in Speech and Gesture

Reference to space involves topological relations, frames of reference, motion events, toponyms and deixis, and has been found to differ across cultures [23]. This paper will focus on the representation of frames of reference in multimodal interaction and will discuss the dynamic interaction with shared cultural knowledge. According to Levinson [7], there are three major frames of reference (FoR) – the object-centred or intrinsic FoR, the egocentric or relative FoR, and the geocentric or absolute FoR. While in an intrinsic FoR a ground object's features are being used for locating a figure, the relative FoR involves the speaker's perspective to create a coordinate system with the help of which both the figure and the ground are located. Finally, in an absolute FoR fixed external features such as cardinal directions are the source of information necessary to locate a figure. Examples (1) – (3) illustrate the expression of different FoRs in speech:

(1) The dog is at the car's front. (intrinsic FoR)
(2) The dog is left to the car. (relative FoR)
(3) The dog is north to the car. (absolute FoR)

While the cross-linguistic differences in the availability of and choice for a certain FoR have usually been investigated concerning speech, Levinson [7, pp. 244ff.] has also listed certain gestural features that may be observed. According to him, gestures within an absolute FoR, such as the one found in Guugu Yimidhirr [24], [13], are characterised by specific

phonological features: the use of extended gesture space, no restrictions towards a dominant articulator, and body torque being used only in those cases where required by biomechanics. Furthermore, gestures and gaze are independent, i.e. a pointing gesture does not necessarily have to be accompanied by eye gaze in the same direction. Also, absolute gestures are often characterised by a certain veracity of the vectors projected by e.g. a pointing gesture – a feature that has been analysed for Guugu Yimidhirr speakers in detail [13]. This means that any vector projected by a pointing gesture directly points to the actual or associated position of a referent. This veracity is constant under rotation and, as Le Guen [6] mentions, includes an absence of metaphorical pointing. If used metaphorically, pointing gestures do not project vectors to the actual position of a referent but rather point into empty space. Further features are the representation of complex vectors in one gesture and the fusion of semiotic types, e.g. iconic and deictic gestures [7]. Finally, gestures have been observed to follow what Levinson [7] calls natural lines – the further away a referent, the higher the pointing gesture. Also, there seems to be a certain distribution of typical handshapes: while flat hand gestures tend to convey information about a vector, locations are more likely to be referred to by index pointing. These characteristics have also been reported for speakers of Yolŋu languages in the Northern Territory of Australia [25]

2. Subjects and Methods

Kreol Seselwa is spoken on the Seychelles, a group of 115 islands located in the Indian Ocean. It is the native language of 99% of the population. Being a creole language, KS is characterised by its mixed nature: while the lexicon is mainly derived from French with occasional influences from other languages such as English or Eastern Bantu languages, its grammatical structure involves creole features such as TMA markers or a fixed S-V-O word order.

The data was collected in 2014 and 2015 on Mahé, the main island of the Seychelles. The data collection involved a triangulation of methods, including a sociolinguistic interview, elicitation tasks and semi-elicited conversations. The interview sessions were conducted with two native speakers who were asked to talk to each other in KS. While the elicitation tasks involved pointing tasks to specific locations on Mahé, route descriptions and the "Man and Tree" space game designed by Levinson et al. [26], the semi-elicited conversations dealt with locally-anchored narrations about the role of family and neighbourhood on Seychelles and a flood that took place in 2013. All interview sessions were video-recorded after the participants explicitly signalled their consent. The data analysed for this paper comprises 67 min of video data coming from 7 native speakers. The data annotation was done with ELAN, and includes detailed annotation of over 700 gesture strokes.

3. Multimodal Reference to Space in KS

3.1. Gestural and Vocal Repertoire for Spatial Reference

As already described in a previous paper [27], KS has several typical word classes at its disposal for expressing spatial reference, such as prepositions, demonstratives and adverbs. For all three FoRs terms are available. The left-right distinction for the intrinsic and relative FoRs can be expressed by the terms *(a) gos / (a) drwat*. Furthermore, the four cardinal directions potentially relevant for an absolute FoR are *(dan) nor / sid / was / les*. The lexical origin of these spatial terms in the French language is clearly visible. This is also the case for toponyms on the Seychelles, such as e.g. *La Misère* or *Beau Vallon*. However, KS is not merely a dialect or 'broken' version of French. Rather, the lexical origins are combined with very idiosyncratic phonological, grammatical and pragmatic structures to form an independent language system. Furthermore, it will become clear in the following sections that the gestural system reveals additional instances of very distinct idiosyncratic realisations and conceptualisations of spatial reference.

Gestural reference to space mainly involves three typical handshapes (see Figure 1): two flat hand gestures (B and 5) and the extended index finger (IX). The two flat handshapes (B and 5) are usually produced in extended gesture space and accompany spatial reference to existing locations beyond the immediate surrounding. Furthermore, they are usually not involved in the specification of a referent. The IX-handshape, on the other hand, is distributed across gesture space and usually accompanies reference to a location in the immediate surroundings, often involving the specification of visible referents.

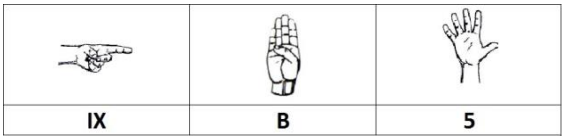


Figure 1. *Handshapes associated with spatial reference in KS.*

The use of gesture space in instances of spatial reference is summarised in Table 1. Over 60% of gestures associated with spatial reference were produced in extended gesture space (periphery or extreme periphery), whereas only 38% of spatially referential gestures were produced in central gesture space (centre or centre-centre). Those gestures produced in the extreme periphery were usually produced in the two upper thirds, i.e. from the waist upwards. Furthermore, they frequently involved back-pointing.

Table 1. *Use of gesture space in KS spatial reference.*

Central gestures (total)	**38%**
Centre	25 %
Centre-Centre	13%
Peripheral gestures (total)	**62%**
Periphery	33%
Extreme periphery	29%

3.2. Dynamic Use of Spatial FoRs

The data analysed suggest that KS speakers apply both a relative and an absolute FoR in their spatial references. As presented in Tables 2 and 3, the two different contexts in which speakers produced spatial reference were the Man and Tree space game on the one hand, and locally-anchored narrations and pointing tasks on the other hand. While in the first context only the relative FoR was used by the speakers, the latter involved a mix of relative and absolute FoRs. This mix of FoRs

has been found to be divided according to modality: the relative FoR is mainly represented in the vocal channel, while the absolute FoR appears almost exclusively in the gestural channel.

Table 2. *Use of FoRs in KS according to context*.

Context	**Frame of Reference**
Locally anchored narrations	absolute-relative
Pointing tasks & route descriptions	absolute-relative
Space game	relative

Table 3. *Representation of FoRs in KS across modalities*.

Frame of Reference	**Modality**
Relative	Speech (Gesture)
Absolute	Gesture

Example (4) illustrates the mix of FoR in a naturally occurring situation. The topic of the conversation was a certain kind of perfume and the speaker's association of it with her former workplace. While in her speech she does not give any spatial information, her gestures show several of Levinson's (2003) characteristics of an absolute FoR. Two subsequent pointing gestures are produced in extended gesture space and are instances of back-pointing. Furthermore, we find a veracity of pointing, as Figure 2 illustrates. In addition, there is no body torque involved and the speaker's gaze does not follow the pointing gesture.

(4) I used to work in a hotel…a hotel de Sesel.

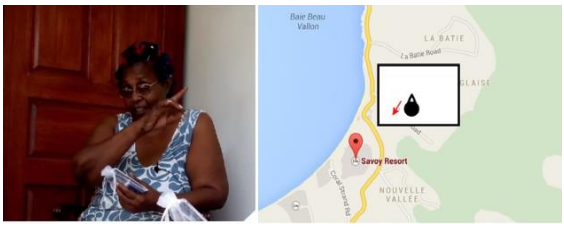


Figure 2. *Features of gestural reference to space in a locally-anchored narration*

In comparison, the gestures produced in the context of the Man and Tree space game did not show any absolute features. Both in the gestural and the vocal channel the relative FoR was applied, as illustrated by Example (5) and Figures 3-4.

(5) En pe vir **anfas ek nou**, enn pe vir **par deryer**.

One is oriented **towards us**, one is oriented **to the back.**

The speaker's viewpoint is made explicit in speech, when the speaker refers to one figure as facing the observer (*nou*).

Figure 3 shows the speaker employing two gestures referring to the orientations of the two figures. Figure 4 displays the differences in both figure orientation (A and B, marked in black) and speaker orientation (marked in red). The left box illustrates this setup at the moment the speaker looked at the stimulus picture. The right box illustrates the setup after the speaker has turned around to describe the picture to her interlocutor. The two arrows represent the vectors projected by the gestures produced by the speaker, which are also shown in Figure 2. It becomes clear that in the gestural channel the original setup between figure orientation and speaker orientation is rotated by nearly 180 degrees. This rotation of not only the speaker but also the conceptualisation of the stimulus setup is exactly what is expected in a relative FoR. In an absolute FoR, in contrast, a change in the speaker's orientation should not have an impact on the orientation of the gestural representation of a figure.

Figure 3. *Gestures produced accompanying spatial reference in the Man and Tree Space Game setting.*

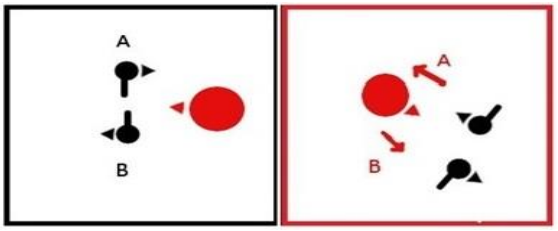


Figure 4. *Features of gestural reference to space in the Man and Tree Space Game setting.*

Further evidence that the gestural system of the participants is not entirely absolute can be found with regard to the different abstractions of pointing gestures. Comparing Levinson's [7] criterion of veracity of pointing and Le Guen's [6] observation of a lack of metaphorical pointing in absolute gesture systems with the data coming from the Seychelles, a mixed picture emerges. Table 4 represents the different levels of abstraction in all pointing gestures analysed, involving not only spatial but also person reference. Metaphorical pointing does occur frequently in the KS references analysed, showing that the KS gesture system also involves features of a relative FoR. Interestingly, however, direct pointing occurs more often in combination with vocal spatial reference than with vocal person reference, whereas metaphorical pointing is more often

associated with vocal person reference than with vocal spatial reference. This suggests that, contrary to what has been described by Haviland [13] for Guugu Yimidhirr, KS speakers may tend to use the absolute FoR only when spatial information is intended to be communicated, and not when person reference is in the focus. However, more data will have to be considered in order to adequately investigate this tendency.

Table 4. *Levels of abstraction in KS pointing gestures*.

Direct Pointing	42%
Metonymic Pointing	26%
Metaphorical Pointing	32%

Finally, the last example will illustrate the dynamic and flexible use of semiotic types in a path description, reflecting the representation of shared cultural knowledge in the KS gesture system. This speaker described a looping path starting and ending in the capital of the Seychelles, Victoria (see Figure 6 for a map of the route described). In her speech, she mainly lists the toponyms of the different locations one would pass by on this path. Her gestures display subsequent and simultaneous combinations of deictic and iconic gestures, as well as her absolute orientation. The gestures displayed in Figure 5 correspond to the path segments 1 and 2. As the pictures illustrate, the speaker's gestures follow natural lines, i.e. the further away the locations indicated, the higher the gesture. At the same time, the speaker iconically treats the locations as something one can 'hold' in ones hands. Furthermore, as evident from Figure 7, the vector projected by the pointing gesture coincides with the actual location of Victoria. During the path description, however, the veracity of the pointing gesture decreases, with only the general direction being represented. Also, the representation of natural lines is being metaphorically extended: instead of representing the distance between the locations and the speaker, the distance travelled is the crucial factor determining the height of the pointing gesture. During the last path segment, represented by the number 3 in Figure 6, the gestures do not include any deictic element anymore. Instead, the speaker represents the path in her gestures by iconically modelling the topographic features of the mountainous path , thus revealing her geographic knowledge of the area (Figure 6).

Figure 5. *Gestures accompanying the path description.*

Figure 6. *Gestures modelling topographic features.*

Figure 7. *Speaker orientation and subsections of the path description.*

4. Discussion

While section 1 has emphasised the interaction between speech, gesture, and culture, section 3 has illustrated that KS speakers employ FoRs in a mixed fashion. The question arising here is why and how a certain FoR and thus a certain strategy of spatial conceptualisation is selected in KS. This question can be addressed best by including the micro-ecology of communication into the analysis. As mentioned in section 1, the micro-ecology of communication includes extra-linguistic factors, such as socio-cultural aspects, which are assumed to interact with the form of communication of a certain speech community.

As mentioned in section 3, whether KS speakers selected the relative FoR only or whether they applied a mix of relative-absolute FoR differed across contexts. One major difference between the two contextual settings is whether shared cultural knowledge, in this case spatial orientation, was available to the speakers as a resource. While in locally-anchored narrations, pointing tasks and route descriptions the speakers referred to existing referents in a concrete environment, the space game included abstract referents in an artificial environment. In other words, the first kind of conversation was characterised by the availability of shared cultural knowledge, whereas the second one involved shared conversational knowledge only. Further evidence that the availability of shared cultural knowledge may influence the selection of a FoR in KS is supplied by the phonological features of the KS gesture system as well as the overall veracity of pointing and the semiotics displayed in gesture in the respective contexts.

One reason why absolute spatial orientation is selected as a resource of anchoring in locally-anchored narration tasks, rather than relying on shared conversational knowledge and a relative FoR, might be the geographic characteristics of the island. Mahé is a rather small island with a clear distinction between coastline and mountainous inland. As a consequence, absolute orientation is facilitated, especially for individuals who grew up and have spent the most part of their life on the island.

In opposition to other communities which apply the absolute FoR to all contextual settings, including small scale descriptions of abstract or fictional referents (see e.g. [13]), KS speakers dynamically switch from one FoR to the other. This hybridity and flexibility reflects other cultural aspects of the micro-ecology of KS communication. Besides the linguistic background described in section 2, mixed heritage can be found in other cultural domains of the Seychelles such as oral traditions, descent, food, or songs. Thus, spatial reference may be regarded as another instantiation of this mixed nature. Furthermore, being a post-colonial society, Seselwa culture may be characterised as a 'third kind', i.e. creatively combining and transforming features of the mixed heritage in an idiosyncratic fashion to form new cultural patterns [29]. Finally, as a language that relies heavily on pragmatic reference marking, KS is characterised by a certain flexibility, especially when it comes to relying on contextual factors. This general flexibility and context-dependency is reflected in the tendency of ad-hoc ascription of different FoRs, as suggested by Pederson [30]. A reflection of this ad-hoc ascription and the flexible switch from one FoR to the other can be seen in the interaction of deictic and iconic gestures produced during the path segment, where the switch between different forms of representation and spatial conceptualisation takes place within one composite utterance.

5. Conclusion

This analysis of KS speakers has shown the interaction of vocal, gestural and cultural factors in the domain of spatial reference. It has been demonstrated that while KS speakers tend to apply the relative FoR in their speech, their gestures display both relative and absolute features, reflecting the availability of shared cultural knowledge. Furthermore, it has been shown that shared cultural knowledge can be represented across modalities. Moreover, it is embedded in a dynamic, context-dependent frame, which relies on the micro-ecology of communication.

As a consequence, this paper highlights the necessity to take contextual, cultural and multimodal information into account in order to achieve a deeper understanding of the processes involved in spatial reference. Further cross-linguistic research combining these three aspects is necessary in order to gain a better understanding of the underlying dynamics of human communication.

6. Acknowledgements

I would like to express my gratitude to all the participants of the study on the Seychelles, as well as Penda Choppy, Cindy Moker, Joelle Perreau and Zan-Klod Mahoune, who supported this project with providing cultural and linguistic insights. Furthermore, I am indebted to Dany Adone for her supervision, support and encouragement. Many thanks go to my colleagues Astrid Gabel and Christina Murmann, as well as the anonymous reviewer of this article. Finally, I would like to thank the following institutions for kindly supporting and/or funding this project: Lenstiti Kreol Enternasyonal, Ministry of Tourism and Culture, University of Seychelles, a.r.t.e.s international, and DAAD.

7. References

[1] D. Efron. *Gesture, Race and Culture*. Den Haag: Mouton and Co. 1972.

[2] A. Kendon. "Contrasts in gesticulation: A Neapolitan and a British speaker compared." in *Körper, Zeichen, Kultur: Bd. 9. The semantics and pragmatics of everyday gestures. Proceedings of the Berlin conference April 1998*. C. Müller & R. Posner, Ed. Berlin: Weidler, 2004, pp. 173-194.

[3] A. Kendon. *Gesture: Visible action as utterance.* Cambridge, New York: Cambridge University Press.2004.

[4] S. Kita. "Cross-cultural variation of speech-accompanying gesture: A review." *Language and Cognitive Processes*, 24, 2, pp. 145–167, 2009.

[5] S.Kita & A. Özyürek.. "What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for and interface representation of spatial thinking and speaking." *Journal of Memory and Language*, 48, pp. 16–32, 2003.

[6] O. Le Guen. "Modes of pointing to existing spaces and the use of frames of reference." *Gesture*, 11, 3, pp. 271–307, 2011.

[7] S. C. Levinson. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge, New York: Cambridge University Press. 2003.

[8] A. Özyürek, S. Kita, S. Allen, A. Brown, R. Furman, R., & T. Ishizuka. "Development of Cross-Linguistic Variation in Speech and Gesture: Motion Events in English and Turkish." *Developmental Psychology*, 44, 4, pp. 1040–1054, 2008.

[9] A. Duranti. *Linguistic Anthropology*. Cambridge. Cambridge University Press. 1997.

[10] C. Geertz. *The Interpretation of Cultures*. New York: Basic Books. 1973.

[11] W.H. Goodenough**.** "Componential Analysis and the Study of Meaning" *Language* 32, pp. 195-216, 1956.

[12] J. Lave. *Cognition in Practice*. Cambridge: CUP. 1988.

[13] J. Haviland. "Anchoring, Iconicity, and Orientation in Guugu Yimithirr Pointing Gestures." *Journal of Linguistic Anthropology*, 3, 1, pp. 3–45, 1993.

[14] D. Wilkins. "Why Pointing With the Index Finger Is Not a Universal (in Sociocultural and Semiotic Terms)." In *Pointing: Where Language, Culture and Cognition Meet*. S. Kita, Ed. Hillsdale, New York: Erlbaum, 2003, pp. 171–215.

[15] D. Adone & E. L. Maypilama,. "Research Report: Bimodal bilingualism in Arnhem Land." *AIATSIS*, 2, pp.101-106, 2014.

[16] A. Nonaka. "Toponyms in Ban Khor Sign Language." *Learning Communities: International Journal of Learning in Social Contexts*, 16, pp. 66–91, 2015.

[17] C. de Vos. *Sign-Spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space: PhD Thesis*. Nijmegen: Radboud University Nijmegen. 2012.

[18] D. Adone & E. L. Maypilama. *A Grammar Sketch of Yolngu Sign Language*. München: LINCOM. 2014.

[19] M. Garde, M. *Culture, interaction and person reference in an Australian language: an ethnography of Bininj Gunwok communication*. Ansterdam / Philadelphia: John Benjamins Pub. Co. 2013.

[20] M. Johnson. *LORE. Capturing Traditional Environmental Knowledge.* Hay River,CA: Dene Cultural Institute and the International Development Research Centre. 1992.

[21] A. Cheikhyhoussef, M. Shapi, K. Matengu, & H. Ashekeleet. "Ethnobotanical study of indigenous knowledge on medicinal plant use by traditional healers in Oshikoto region, Namibia." *Journal of Ethnobiology and Ethnomedicine*, 7, 10, pp. 1–11, 2011.

[22] M. Durie. "Understanding health and illness: research at the interface between science and indigenous knowledge." *International Journal of Epidemiology*, 33, 5, pp. 1138–1143, 2004.

[23] S. C. Levinson & D. P. Wilkins. *Grammars of Space: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press. 2006.

[24] S. C. Levinson. *Language and cognition: the cognitive consequences of spatial description in Guugu Yimithirr. (Working*

Paper #13). Nijmegen: Cognitive Anthropology Research Group at the Max Planck Institute for Psycholinguistics. 1992.
[25] D. Adone. personal communication, 2014.
[26] S. C. Levinson, P. Brown, E. Danziger, E., L. de León, J. Haviland, E. Pederson & G. Senft, G. "Man and Tree & Space Games". In *Space stimuli kit 1.2* S.C. Levinson, Ed., Nijmegen: Max Planck Institute for Psycholinguistics, 1992, pp. 7–14.
[27] M. A. Brück. "Bimodal reference marking in Kreol Seselwa." *Island Studies*, 2, pp. 20–26. 2015.
[28] D. McNeill. *Hand and Mind.* Chicago, London: The University of Chicago Press. 1992.
[29] H. K. Bhabha. *The Location of Culture*. London, Routledge. 1994.
[30] E. Pederson. "How Many Reference Frames?" in *Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence: Vol. 2685. Spatial Cognition III. Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning* C. Freska, W. Brauer, & K. F. Wender, Eds. Berlin, Heidelberg: Springer, 2003, pp. 287–304.

Zero Mean Lag Communication Over Networks: A Route to Co-Presence?

Fred Cummins[1], *Jonathan Byrne*[1]

[1]University College Dublin

fred.cummins@ucd.ie, jonathan.byrne@ucd.ie

Abstract

We contrast two ways of thinking about communication: communication as message passing, and communication as reciprocal coordination. From the invention of writing to the ubiquity of SMS, speech and language technology has uniformly employed the first model, and thereby done nothing to support, extend, or explore the second model. We suggest that the coordinative approach is better suited to understanding how face to face interactants establish co-presence. The technical challenges of establishing co-presence amounts to achieving synchronisation with a mean lag of 0 ms. We suggest that this goal might be approached through the exploitation of predictive models for behaviours that are inherently constrained, or known to both parties. Although we have not yet succeeded in achieving this goal, we chart a possible route of future exploration, with the distal goal of allowing people to engage in strongly synchronised behaviours such as chanting over networks.

Index Terms: co-presence, reciprocal interaction, liveness

1. Introduction

Communication can be thought of in more than one way. If we view communication as *coordination*, we are emphasising the manner in which your activity and mine become non-independent. If we view it as *message passing*, we focus instead on how my ideas and thoughts can be transferred to you through some medium. The first, coordinative, view emphasises the reciprocity of communication: you affect me and I affect you, simultaneously, and the distance between us is lessened. The second, conduit, view foregrounds the content of the exchange, but portrays the participants as relatively independent.

The coordinative perspective is most clearly demonstrated through the modality of touch. This is the primal mode through which infants first experience bonding with their mother, and it remains the most intimate form of communication, so much so that we ration and regulate who is allowed to touch whom, where, and when. When we touch, there is a two-way continuous connection between us that does not admit of dissection into independent components [1]. The two hands in a handshake cannot properly be understood as separate. In touch, the two subjects are necessarily *co-present* in a very important sense. The reciprocity of communicative touch is the reason why we do not have a recording medium for touch, analogous to our use of pictures or recorded sound. Nevertheless, co-presence is not a merely haptic phenomenon. In each other's presence, we clearly and mutually influence each other's gaze, the sounds we make, the manner in which we move and so on.

This contrasts starkly with communication conceived of as message passing, which has been often taken to be the essence of linguistic communication. Considered in this fashion, communication involves my private intentions and thoughts acquiring some kind of encoding—in sounds, text, images—and then being transferred to you for decoding. The communicating partners are here treated as separate entities and the act of communication is directed: either from me to you, or vice versa, but the act is essentially decomposable into sender and recipient.

This is not to insist that there are two categorically different kinds of communication, but to point out that we can describe, attend to, model, and perhaps facilitate coordinative or message-passing aspects of a given communicative situation. Remarkably, the technological support of speech and language has taken as its object the support of message-passing aspects of communication alone. From the development of writing—a technological breakthrough that set in motion a series of profound cognitive changes—to the most recent forms of messaging through smart phones, communication has been seen as one thing only, and the very possibility of augmenting or facilitating the coordinative aspects to communication has been overlooked [2]. It is to this that we turn our attention.

In what follows we seek to articulate the problem of establishing presence when interactants are only in touch (as it were) over networks. We do not yet have a working implementation of a communicative protocol that can support a sense of co-presence, but we hope it might be of some benefit to lay out the territory, sketch the logical form of a possible solution, and to show how initial exploration of this novel space of technological innovation might proceed.

2. Language and Joint Speaking

The development of many kinds of technological support for message passing has gone hand in hand with a specific view of the nature of language that has occupied the core of the discipline of linguistics for over a hundred years. We might characterise the first half of the 20th Century as the structural era, with the works of Saussure playing a central role, while the second half clearly belongs to the generative grammarians, whose most visible figure was Chomsky. Saussure formalised the study of *langue* over *parole*, thereby emphasising the abstract, systematic, formal aspects of linguistic communication, and self-consciously stepping over the messiness of verbal behaviour (a limitation of which he was painfully aware). Chomsky likewise valorised *competence* over *performance*, seeking to characterise an abstract underlying system that was at some remove from the messy business of speaking. Both programmes viewed speech as just one mode through which language finds expression, and language was understood as the exchange of propositional content encoded in rule-governed sequences.

This view of language has an odd historical flavour to it. After all, however we characterise language, it is surely as something which arose uniquely in our species and which differentiates us sharply from our nearest cousins, the great apes. To emphasise the symbolic, mode-agnostic, characteristics of some forms of language use is to adopt a perspective that doesn't care

about the differences between speech and writing. Yet writing has only been around for about 5,000 years, widespread literacy for no more than 500 years. Whatever happened to our species that gave rise to society, culture and human intellectual life is very much older than this, and the voice has been its primary vehicle all along.

There is a common form of speaking, found in every culture on Earth, and central to the affairs of all societies, that is ignored when we view language in this abstract, intellectualised way. This is joint speaking, which is found whenever multiple people say the same thing at the same time [3]. This is the form of speech found in all major religious traditions, frequently built into rituals which play an axiomatic role in establishing a common order. It is also found in situations of protest, when collectives give common voice to common concerns. And it is the mode of speech in which football fans enact a common identity on the terraces. As different as these domains are, joint speech displays some superficial characteristics that transcend the domains and that speak eloquently of the collective subject: The absence of any differentiation between speaker and listener stands in marked contrast to the abstract message passing view of language, as everybody is both speaker and listener, and everybody already knows the text. We also find a continuum of prosodic forms, with no clear distinction between speech and music, the English word "chant" serving double duty in both worlds. We find a central role of repetition, which makes sense only if we acknowledge the performative nature of joint speech: whatever is being achieved through this practice, it is achieved in real time only and through the urgent participation of all concerned.

In joint speaking, a highly charged form of co-presence is brought into being among interactants in a manner entirely unlike a sequenced exchange of messages [4]. Most people have experienced the sense of loss of personal autonomy (or its transference to a group) when taking part in chanting during protest or in support of a team. Choral singers are familiar with the remarkable sense of experiential blending or transcendence that arises when singing in unison. But perhaps the most eloquent illustration of the power of the co-presence that arises in this fashion is given, not by saying anything, but by being silent together. To be silent on one's own achieves nothing, but to be silent collectively, in joint commemoration of tragedy, illustrates viscerally the power of co-presence and the relative unimportance of the lexical content. The frenzy of an angry mob, or the ecstasy of the heavily embodied chant of Sufi *dhikr* likewise reveal the power of joint speech for turning collective activity into something highly charged, something shared, and something that is enacted by doing [5].

The focus on language as message-passing has led to many technologies that allow us to encode and transmit messages of many forms. But there are no technologies that allow us to speak together. This seems odd, and it is not hard to think of uses for such technologies. In what follows, we will first illustrate the problem, and then go on to discuss how the next steps in illuminating this largely unexplored space of potential innovation might proceed.

3. Liveness and Skype

Users of Skype or similar services are frequently aware that there is a slightly unreal feel to the conversation. Some of this, particularly in older implementations, is due to the mismatch between the spatial location of the camera and the position of the eyes of the interlocutor. There is also a small temporal lag, but neither of these two factors is typically an obstacle to carrying on a conversation. VoIP services were developed for the purposes of conversation, and the current standards seek to guarantee a lag of no more than 150 ms end to end [6]. Under these circumstances, we can take turns in a conversation, but we can't chant. If two people try to sing or speak in unison, this seemingly small delay is compounded, leading to an inevitable breakdown in the coordination necessary to synchronise.

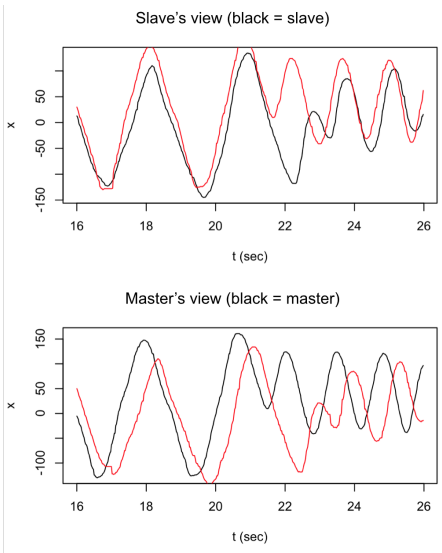


Figure 1: *Time (X) versus horizontal position (Y) for oscillatory finger movements produced in master-slave mode.*

The problem is illustrated in Fig. 1 which shows horizontal finger position against time for a networked application (described below) in which subjects at each end of a network make manual oscillatory movements that are displayed, with a 150 ms lag, on the screen of the other. In the example shown, the lower trace belongs to the "master" who was told to ignore the movements of the other, "slave", participant. The slave, on the other hand, was told to synchronise with the master. While the slave succeeds quite well, the master experiences two traces that are displaced in time by approximately 300 ms. Under these conditions, the equitable and reciprocal form of co-regulation that is possible in the flesh, is rendered impossible.

We frequently speak of "liveness" as if it were clear whether something is live or not. In a world filled with recordings, and edited creations, the distinction seems fairly straightforward. But with a little thought we can show that liveness is not an all-or-nothing affair. Consider a concert in which Jools Holland plays a duet with you. If you make a mistake, that will adversely affect the joint performance. You are very intimately involved with one another in live interaction. Now consider the same concert, but this time you are in the "live" studio audience. You have a sense of co-presence, and you could, at a push, influence things, e.g. by shouting or throwing something, but the manner in which you are involved in the whole scenario is rather different, and your role much smaller, than when playing the duet with Jools. Shift scene, and you are now at home, watching a "live" broadcast of the show. The advertised "liveness" does matter. You have a sense that things could go wrong, the unexpected could happen, and the spectacle is thus somewhat fragile.

But your capacity to influence things is now minimal. Change things just slightly, and you are watching the "live" performance with an hour's, or a year's, delay. There is a meaningful sense in which it is still live: the action is unbroken, the performance is probably less polished and somewhat less predictable than a studio recording, but the manner in which you are involved has now been watered down to homeopathic proportions. Liveness, then, admits of a good deal of variation, but it finds its strongest exemplar when several people are physically co-present to each other, deeply involved in each others' actions.

At first blush, it would seem that this is simply how things must be. If we communicate over networks, there must be a lag, because transmission times are non-zero, always. Synchronisation would seem to demand a zero-mean lag, and this sets an engineering goal that is unreachable in principle. We may be able to reduce lags well below 150 ms (and it is possible to perform a reasonable chant over landlines, as opposed to VoIP), but we cannot eliminate them. But we believe that this kind of thinking is, itself, beholden to the singular view of communication as message passing, and if we adopt instead a coordinative view, an unexplored opportunity for technical exploration opens up.

4. The Mirror Game

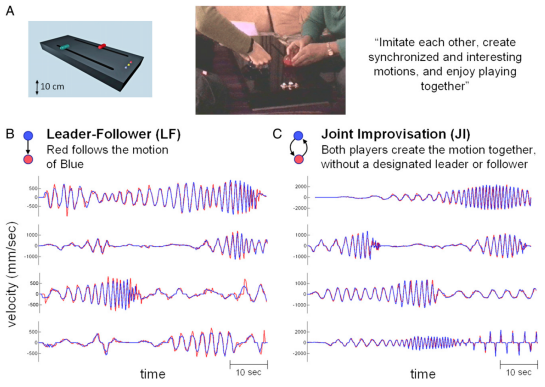


Figure 2: One-dimensional mirror game used in Noy et al. (2011). A: Movement of the sliders is sampled at 50 Hz. B: Sample velocity traces from a leader-follower round. C: Sample velocity traces from a joint improvisation round. From Noy et al. (2011)

There is a form of collective exercise known as the Mirror Game with origins in improvisation theatre and movement therapy [7]. In this, two or more participants improvise sequences of movements in one of two modes. In leader-follower (LF) mode, one person dictates the movement sequence while the others try to follow. In the second mode of joint improvisation (JI), nobody has the assigned role of leader, and synchronised activity must emerge spontaneously. Noy et al. created an experimental variant of this game in which movement is confined to the horizontal movement of a slider [8]. They found that patterns created in the two experimental conditions were equally complex, but that synchronisation was, on average, somewhat better in the JI condition. In particular, in JI rounds they would sometimes find periods of co-confident motion in which the two horizontal traces remained in lock step with no appreciable jitter (Fig: 3). A follow up study [9] they found that such co-confident motion displayed curvature that was qualitatively different from motion in which one player dominated or led. In [8], the authors introduced a reactive-predictive control model in which each participant used a simple model to predict future trajectories of the other.

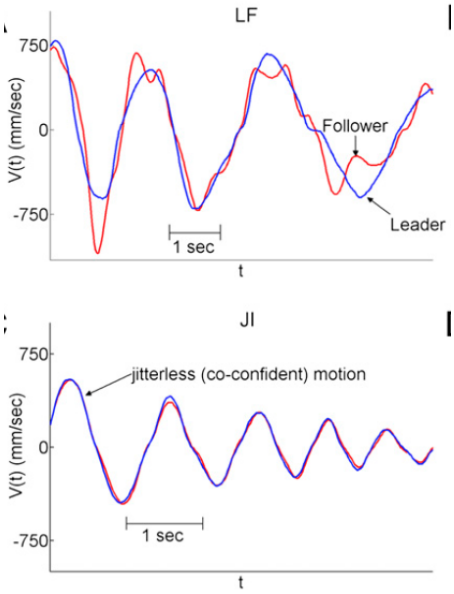


Figure 3: Example of co-confident motion (bottom) contrasting with more jittery trajectories found in LF-mode. From Noy et al. (2011)

No networks are involved here, of course. Players are synchronising in real face-to-face live interaction. However, the signature of co-confident motion illustrated in Fig. 3 may provide an important target that a comparable setup over networks might aim for. The mirror game thus provides us with a potential empirical index of reciprocal coupling among actors.

5. Towards zero mean lag over networks

In order to see how one might approach zero mean lag over networks it is useful to recall again the fundamental difference between communication as message passing and communication as coordination. In the first case, there is uncertainty about the message that will be transmitted. One must therefore wait until a signal has been received before one can know what it is. In the latter case, we are frequently dealing with a situation in which all parties know the sequence of movements, or words, that are to be performed. If that is the case, then for each subject, the future actions of the partner are largely predictable from the past. We can exploit this predictability in the behaviour of the interacting parties.

Let us consider two exemplary implementations of a notional zero-mean-lag system: long-distance chanting, and remote playing of the mirror game. In the case of chanting, predicability arises because the same text is repeated many times over. If the text is not known at the start, then it is available after a single iteration. Chanting thus is inherently predictable. In the case of the mirror game, trajectories along the rail are greatly constrained by physical contingencies. They will be

continuous, and at any given moment, knowledge of the position and velocity of the marker at the last several time steps, $t-1, t-2\ldots$ will allow confident prediction, within some margin of error related to the step size of discretisation.

Assuming that the signal can be discretised into equally spaced samples, and assuming a minimum transmission lag of one time step at each point, we simply ensure that what one person hears/sees/encounters at any time is the best possible prediction, based on recent values of the incoming signal. In what follows, $A(t)$ refers to the signal generated by person A at time t, and $\hat{B}(t)$ to the prediction of where B is most likely to be at time t. For simplicity, we assume prediction based on a single previous time step, but the approach should generalise to a sliding window of prediction.

At time t,
$\ldots A$ sees/hears $\hat{B}(t)$
$\ldots A$ says/does $A(t)$
$\ldots A$ receives $B(t-1)$
$\ldots A$ updates predictive model based on $B(t-1)$

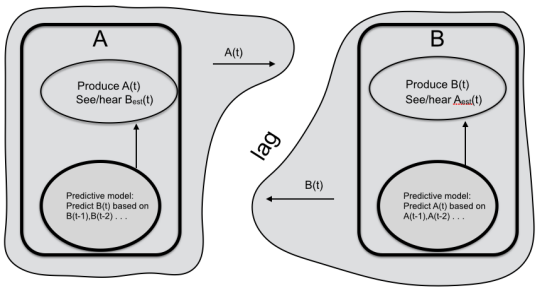


Figure 4: Basic architecture for exploiting prediction. Although there are lags in transmission from A to B, at any moment in time, A hears/sees/encounters a signal without lag, with fidelity that is proportional to the predictability of the signal.

This way of framing the problem and its solution is wholly generic. As long as the behaviour of one system is, to some extent, predictable by the other, a solution of this form can be implemented. The confidence with which the value of the signal at t can be predicted based on the last n values seen will be an important determinant of the strength of the mutual entrainment that can be achieved.

We are implementing a pilot system for exploring this idea (Fig. 5). We use two networked computers equipped with Leap Motion controllers [10]. In our initial implementation, each player sees a marker on screen corresponding to their own hand position in the vertically-oriented X-Y plane, along with another marker corresponding to a *prediction* of the other person's position, based on a sliding window of previous observations. Biological motion is constrained by the requirement that it be continuous, that motion be physically plausible, etc, so that the position of the co-player's hand can be predicted. For testing purposes, we include the option of introducing a specific fixed lag between packets exchanged over the network.

Fig. 6 illustrates the situation without any prediction. It shows asynchrony for trials run in master-slave mode, where the master ignores the slave, while the slave tries faithfully to synchronise with the master. As the transmission lag increases, so the asymmetry between the views of the two subjects becomes more extreme. The slave manages fairly constant performance across conditions, while the master becomes increasingly asynchronous with respect to the slave.

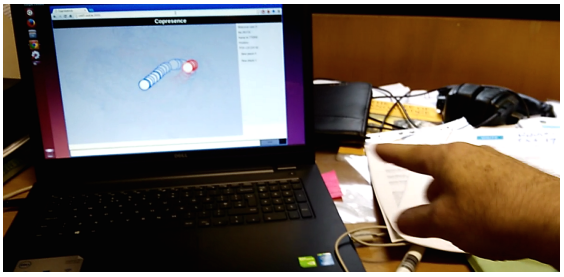


Figure 5: Snapshot of a pilot system under development

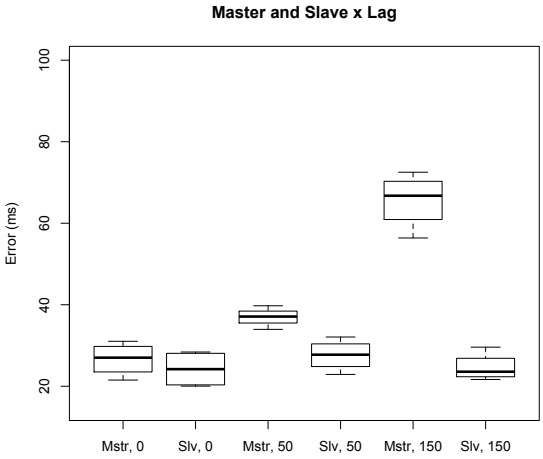


Figure 6: Measured asynchrony for both master and slave where the master ignores the slave, while slave synchronises with master, with no predictive model. Lags of 0, 50 ms and 150 ms were used.

We have tried several predictors so far, including a weighted linear, and a polynomial spline. At this point, however, a performance increase over the baseline has not yet been achieved. The system is intended to act as a crucible for refining questions about how a predictive system might overcome the challenge of zero mean lag synchronisation. In the remainder of this article, therefore, we consider the challenges ahead.

6. Extension to speech

If coordinated finger wiggling over networks, is challenging, one might opine that synchronising speech is even more so. We believe there are grounds for cautious optimism, however. One important aspect to chanting is to note that the text of the speech or song is known to both parties, and so does not, itself, need to be transmitted. Rather, because the sequence is known, the participants need only to keep track of where they each are in the sequential unfolding of that sequence over time.

A standard linear predictive encoding of speech produces a single vector for each windowed frame of the speech signal, and

good intelligibility can be obtained with a transmission rate of about 50 frames per second. This is a relatively sparse encoding, but still poses a significant challenge for prediction. For a known text, however, each participant can have a model sequence of LPC vectors, so that prediction becomes the easier task of estimating where, within the known sequence, the other participant is now. The known reference sequence could be generated using text to speech models, or, in an iterative process, the previous actual enunciation of the standard text might be employed. This remains as territory to be explored.

7. Conclusions

We have attempted to outline what a future technology of co-presence would be like. It would allow remote participants to establish genuine synchrony for known, or predictable, patterns of behaviour, including both hand movements and speech. It is our understanding that rich reciprocal coupling is a necessary prerequisite for engendering a sense of co-presence, and that this can be approached through the restricted domain of hand movement, and then extended to joint speech. The space of whole body synchronisation might be approached once progress has been made in these rather more restricted domains. Despite the initial difficulties, we believe that the development of prototypes and pilot systems may be a good way of teasing out the complexities of the field, and may be the starting point for a genuinely different form of communicative technology.

We foresee potential application in domains of human activity that have not yet benefitted from technological support, and we would suggest that one such domain is the participation in rituals, which frequently demands synchronisation of both speech and gesture. Participation in rituals is an important means by which cultural and religious identities of various kinds are maintained, and the enormous numbers of migrants, both voluntary and involuntary, suggests that there is a very large potential user base for any such application.

8. Acknowledgements

The prototype implementation is supported by a UCD seed funding grant to the first author.

9. References

[1] M. Ratcliffe, "Touch and situatedness," *International Journal of Philosophical Studies*, vol. 16, no. 3, pp. 299–322, 2008.

[2] W. J. Ong, *Orality and literacy*. Methuen & Co. Ltd., 1982.

[3] F. Cummins, "The remarkable unremarkableness of joint speech," in *Proceedings of the 10th International Seminar on Speech Production*, 2014, pp. 73–77.

[4] ——, "Voice, (inter-) subjectivity, and real time recurrent interaction," *Frontiers in Psychology*, vol. 5, 2014.

[5] ——, "Towards an enactive account of action: speaking and joint speaking as exemplary domains," *Adaptive Behavior*, vol. 21, no. 3, pp. 178–186, 2013.

[6] *Recommendation ITU-T G.114 One-Way Transmission Time*, Int'l Telecommunication Union Std., 1996.

[7] R. Schechner, *Environmental Theater*. New York: Applause Theatre and Cinema Books, 1994.

[8] L. Noy, E. Dekel, and U. Alon, "The mirror game as a paradigm for studying the dynamics of two people improvising motion together," *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 20 947–20 952, 2011.

[9] Y. Hart, L. Noy, R. Feniger-Schaal, A. E. Mayo, and U. Alon, "Individuality and togetherness in joint improvised motion," *PloS one*, vol. 9, no. 2, p. e87213, 2014.

[10] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[11] J. Laroche, A. M. Berardi, and E. Brangier, "Embodiment of intersubjective time: relational dynamics as attractors in the temporal coordination of interpersonal behaviors and experiences," *Frontiers in psychology*, vol. 5, 2014.

The Role of Eye Gaze and Body Movements in Turn-Taking during a Contemporary Dance Improvisation

Vito Evola, Joanna Skubisz, and Carla Fernandes

BlackBox Project, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa
{vito.evola, joanna.skubisz, carla.fernandes}@fcsh.unl.pt

Abstract

This paper intends to contribute to the multimodal turn-taking literature by presenting data collected in an improvisation session in the context of the performing arts and its quali-quantitative analysis, where the focus is on how gaze and the full body participate in the interaction. Five expert performers joined Portuguese contemporary choreographer, João Fiadeiro, in practicing his Real Time Composition Method during an improvisation session, which was recorded and annotated for this study. A micro-analysis of portions of the session was conducted using ELAN. We found that intersubjectivity was avoided during this performance, both in the performers' bodily movements and mutual gaze; we extrapolate that peripheral vision was chiefly deployed as a regulating strategy by these experts to coordinate turn-taking. A macro-analysis comparing the data with an analogous one obtained from Non-Performers provides the context for a discussion on multimodality and decision-making.

Index Terms: gaze, non-verbal behavior, silent turn-taking, gesture function, decision-making, performing arts, inter-rater agreement

1. Introduction

Humans regulate their contributions in social interactions using practices, norms, and rules depending on the nature of their exchanges (inter alia [1]), whether it be by using prosody to solicit a reply to a question or realizing who goes next around the table during a hand of poker. The present study intends to contribute to the multimodal turn-taking literature by presenting data collected in a group contemporary dance improvisation where speech is absent. The quali-quantitative analysis presents preliminary results of how body movements alone (i.e. without the support of language) have the onerous of communicating and coordinating in the interaction.

For the purpose of this study, five expert performers joined internationally renown Portuguese contemporary choreographer, João Fiadeiro, in practicing his Real Time Composition (RTC) Method (or *Composição em Tempo Real*; [2]). Fiadeiro, one of the founders of the *Nova Dança Portuguesa* in the 1980s, created the so-called "RTC Game" in 1995 as an improvisation exercise in order to provide choreographers and performers a methodological tool for composing artistic works.

Applying the method, the artists take turns performing in a delimited space in the studio, following a process of creating relations with previous actions in the piece. Although Fiadero's method invites performers to use their bodies on a stage floor, he also uses a variation using props on a table. As the performers sit around the table, and through means of self-selection, they perform a single action at a time on the Game Table with props taken from the Objects Table to develop compositions. This improvisational performance is called a "Game". Creative and innovative ideas and material for stage compositions and other types of performances are generated collaboratively through what emerges throughout the Game.

Unlike other research done on "expressive gestures" in the domain of dance (inter alia [3]), which focuses on those non-verbal behaviors having an affective content (concerning the performer's persona's mood, feelings, emotions, etc.) which performers have rehearsed and act out on stage, we are more interested in the behavior which is less monitored and not explicitly intended for an audience. The focus of this study is more on the "behind the scenes" behavior, concentrating on those moments where expert performers are not performing per se, but have to make decisions of what, how, and if to perform next and at which moment during an improvisation, and all in coordination with their fellow performers' behavior.

In contrast to previous studies on turn-taking in social interactions, the context of this inquiry is linguistically independent, and there are no regulated turns in the traditional sense. Performers do not talk to each other during the improvisation unless their speech is being used as artistic material. They are also free to choose to perform in the improvisation or not, but only a single action at a time, and not twice in a row. Nonetheless, there is social communication: turns are coordinated by the information "given" (e.g. moving towards the table to perform) and information "given off" (e.g. via gaze or other body movements) [4].

Various studies investigating the co-occurrence of speech and gestures in the turn-taking scenario confirm that interlocutors systematically use their non-verbal behavior to coordinate the conversational flow. The gesture involvements in the regulative process of turn-taking mechanism was sufficiently examined in previous research in multi-party conversations [5, 6], but mainly in dyadic situations [7, 8], suggesting that people deploy a broad scope of body movements to yield or grab the floor (e.g. pointing gestures [9], head movements [10, 11], eye gaze [12, 13, 14], and body posture [15, 16]).

To our knowledge, ours is the first study to address the issue of turn-taking where speech is accessible but not used. This study intends to describe and analyze what non-verbal strate-

The Methods part (2), section 4.3, part 6, and all tables and figures were contributed by the second author and revised and rewritten together with the first author. The Qualitative Analysis section (3.2) was written by the third author. Study design and implementation, and the remainder of the paper is the work of the first author, revised on the basis of input from the second author. The creation of the annotation scheme and the data processing was shared between the first two authors.

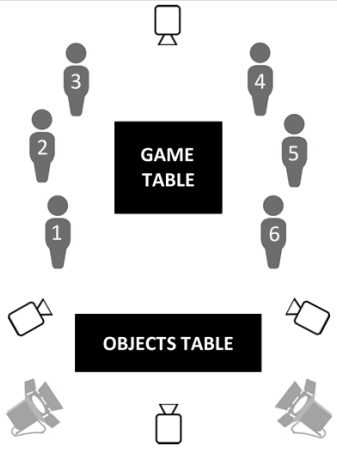


Figure 1: Schematic representation of the setup.

gies are deployed in coordinating complex turn-taking actions in a creative multiparty social interaction where speech is not involved, and whether or not these strategies are analogous to other social interactions where speech is co-present. We will compare what has been described in the literature with the analysis of empirical data to address the question of how performers coordinate their bodies differently by looking at the cues they "give" and "give off" when the speech channel is not used as a communicative tool. We will also describe qualitatively the role of decision-making in the improvisational performance.

2. Methods

2.1. Participants

Five Expert performers who were also practitioners of RTC participated in the study alongside the choreographer Fiadeiro for a group total of six. All participants had at least eight years of professional dance/performance formal training and experience, and on average three years of formal training and experience in Fiadeiro's method. The group was balanced for gender (3 females and 3 males) and culturally mixed (Portuguese and non-Portuguese). Participants were between the ages of 26-41 and proficient in English, which was the common working language, although the Game performances were silent.

2.2. Materials and Procedure

The study was conducted at the RE.AL Atelier, Fiadeiro's studio in Lisbon. The six participants were seated about 1.5m away from and around the Game Table, the focal point of the Game performance. Props to be used during the improvisation exercise were readily available on the Objects Table (Fig. 1).

The study was conducted in a 3.5-hour session, which included briefing, debriefing and breaks between improvisation exercises. After having been informed and given their consent, participants were briefed and had 2.5 hours to perform various Games.

2.3. Data and annotation

The collected video data of the Game performance, excluding briefing, debriefing and breaks, totals 51 minutes. Given the lack of resources to annotate the entire data, a sampling destined for micro-analysis of at least the first 10% of the Game performance was decided a priori. This data subset of the first six minutes of Expert Game performance was processed, annotating the movements of each of the six participants.

An annotation scheme was created in-house for the purpose of this study. In order to investigate how the performers interpret their co-performers' bodily signals, and thus anticipate the communicative flow in their current social environment, the focus of the annotation scheme was on the their visible behavior, as it was perceived and interpreted by those sharing that particular context. This including all movements, gross or fine, that other participants potentially noticed and as they were captured by the four HD cameras.

The annotation scheme codes for information related to:

1. directedness behavior (spatial location and orientation of the body, gaze points, object interaction);
2. a formal description of each movement unit, or MU (i.e. a gestural complex marked by the distinct change of the articulator's configuration or position in space) of the head/face, upper body, and lower body articulators. Each annotation was comprised of the temporal segmentation (defined by each MU's onset and offset) and a label indicating the articulator(s) formally and actively involved;
3. a hermeneutic tier categorizing the functional-semiotic interpretation of the MUs. Each MU temporal segmentation was labeled according to one of the three functions below. This functional taxonomy is a semiotic classification, based on Peircean relations of firstness, secondness, and thirdness [17]. This follows a hierarchical taxonomy where the higher order builds on (and includes) the lower one(s):
   (a) self-focused MUs (purely physical movements meant for the self; e.g. fidgeting, stretching);
   (b) context-focused MUs (relational movements establishing a physical or cognitive relation of orientation, attention, volition, etc.; e.g. orientational head-turns, action-oriented movements, deliberate attention-gaining actions, etc.);
   (c) communication-focused MUs (representational movements, having a symbolic or social nature; e.g. polite smiles, symbolic-communicative gazes, etc.).

Semiotically speaking, any body movement could be inferred and construed as communicating something, even when the person does not have the intention of communicating that. For example, we may infer that a person fidgeting is nervous, but that would not necessarily make that action communication-focused (unless it was done deliberately to convey that sense). Our analysis is obviously not from the production perspective: we do not have access to what the participants were thinking while they were performing these body movements. We can, however, speculate that from an interpretative perspective a certain movement was intended to be communicative (in a strict sense) or not, notwithstanding what information can be inferred.

The first two levels of annotation have a more objective quality (for example, the participant is seated or is moving to a new location; there was movement or not in the left hand, right leg, head, etc.), whereas the last level, based on the previous formal MU segmentations, describes raters' subjective interpretation of the performers' movements before, during and after

Table 1: The global results of the inter-rater agreement obtained from the modified Cohen's kappa [18] calculated in ELAN. The measurement was conducted using data from three participants of the 6-minute subset.

	Global results		
Participant	*kappa*	*kappa_max*	*raw_agreement*
P4	0.6352	0.7590	0.6500
P5	0.9041	0.9281	0.9111
P6	0.9516	0.9516	0.9541

Table 2: Distribution of the functions of all movement units, all participants.

		Function		
	n	self-focused	context-focused	comm.-focused
head/face	141	76	65	0
upper body	122	109	13	0
lower body	58	57	1	0
∑	321	242	79	0

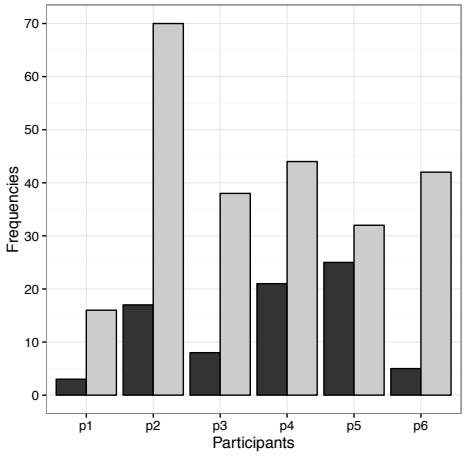


Figure 2: The relative frequencies of functions detected in the head/face, lower and upper body regions of the six Expert participants. Only context-focused (dark gray) and self-focused (light gray) movements were present in the data (zero instances of communication-focused MUs).

their actions. According to high inter-rater agreement (see below), we extrapolate that co-participants who attended to these same movements interpreted each other's behavior in a similar way.

2.4. Inter-rater agreement and data reliability

Because of the importance of data validity and reliability in any research endeavor, working with a reliable annotation scheme was crucial for this study. For this, the scheme was gradually improved upon and eventually tested on the data collected from pilot studies. Two annotators processed a sample from the first pilot study using the annotation scheme, which was critically discussed and reviewed.

The revised version of the scheme was then applied to a sample from the second pilot data for validation. The result of the modified Cohen's kappa [18], calculated in ELAN, produced a global agreement of κ= 0.8685, considered an "almost perfect agreement" [19]. This value confirmed the validity of the annotation scheme, which was then used on the final data.

Two raters annotated three of the six participants, 50% of the data, using the final version of the annotation scheme. Based on the kappa obtained from this sample (Tab. 1), one annotator proceeded confidently with the coding of the remaining 50% of the data.

3. Results

3.1. Quantitative results

The data from the Expert Game performance yielded a total of 1186 annotations. The comparison of all movement unit (MU) functions across participants indicates variety across individuals (Fig. 2); nonetheless, trends in the data do emerge, and some generalizations can be made with regards to both the body movement data and the gaze data. In particular, unlike in more common, everyday social interactions, we found that intersubjectivity was avoided during this performance of contemporary dance improvisation. These results will be discussed below.

3.1.1. Gesture and body-movement function data

The data indicates that the participants performed three times as many self-focused movements (*n*= 242) than context-focused movements (*n*= 79; see Tab. 2). About half of these self-focused movements were produced in the upper body; in fact all but 13 MUs performed with the upper body were self-focused. One out of every three self-focused MUs were produced with fingers and one or both hands (*n*= 58; 33.5%).

Context-focused movements were a third fewer, present mainly in the head/face region, clearly because of changes in head orientation between the two tables, which were the two main focal points throughout the exercise.

Zero (*n*= 0) communication-focused movements were found in the data.

3.1.2. Gaze data

The usage of the term "gaze" in this study may be better described as gaze direction or the end-point of the gaze (to other participants, to the table, etc.). The method of gaze analysis adopted is purely based on what the annotators perceptually coded in a frame-by-frame video analysis and does not include saccades and other minor movements, which may require eye-tracking devices to collect fine-grain information. This technique has been successfully adopted in previous gesture and gaze research (inter alia [7, 20]).

The data exhibits few and fleeting moments of gaze contact among participants. Considering the group as a whole and the time all participants spent looking at another participant versus looking elsewhere, each participant spent on average slightly over 10 seconds each minute glancing and looking to any other participant. Mutual eye contact, where any two participants look at each other reciprocally, amounts to 1.1 seconds in the entire 6 minutes distributed over three distinct occurrences, two of these mutual glances taking place within the first 30 seconds of the data when the Game was just underway.

3.2. Qualitative analysis

We present a preliminary qualitative macro-analysis comparing the 3.5 hour session of the Expert performers' Game interactions with analogous data collected from a parallel study involving a group of Non-Performers. This synoptic analysis focuses on features directly related to the decision-making process throughout the improvisation exercises, such as the management of turns and hesitations versus determination movements in both groups (when participants are moving from their chairs to the Objects Table). We were closely looking at torso and arm movements: determinedly leaning forward just once before standing up and/or backwards in the chairs when there is any hesitation. The differences between the two groups have been analyzed under the light of recent literature focusing on social cognition and decision-making [21]. Constraints such as common knowledge [22], alignment [23, 24] and trust [25] have been taken into account to contrast the results between groups.

Regarding how turns were managed, the Experts took much more time in between turns as opposed to the other group of Non-Performers. Turn management was much more fluid compared to the Non-Performers group, which was not as confident with the method. Experts have been trained to concentrate, taking their time before acting, and to focus only on the Game Table as they are quite used to performing the Game with Fiadeiro: there is somehow a similarity to meditation practices, where silence and control over body movements seem to rule.

As expected, Non-Performers rely much more on those strategies common in verbal social communication, such as gaze exchange. We observed many more gazes to each other and to Fiadeiro, probably looking for confirmation before acting. Moreover, their posture sitting in the chairs seems to be quite stiff.

Concerning determination versus hesitation differences between both groups, the Experts did hesitate much less before taking action than the Non-Performers, which was not surprising either, due to their very different levels of acquaintance with the Game. The Non-Performers' higher hesitation rate can also be related to the fact that they perceive the choreographers' presence as an "authority", whom they are implicitly hoping to please by following his example. They seem to need his approval and reassurance by looking at him before taking action. Another possible reason for their hesitations (either by moving on their chairs uneasily or by looking at the other participants or Fiadeiro before deciding to stand up) can be their tendency to compete with each other by trying to be the "best pupil" in the eyes of the choreographer who does not know them yet [26, 27, 28].

The Expert performers seem to be very relaxed and focused, almost as if they were meditating and reading each other's minds (mentalizing). It seems that they have developed higher-level control processes which modulate low-level reactions such as emotional impulses. Moreover, because they have been playing the Game with Fiadeiro for many years, they perceive him as one of them, not there to judge but to collaborate with (to simply play with the collective intention of creating "common ground", in Fiadeiro's words). According to [22], when acting collaboratively, each subject may automatically represent the task requirements and goals of the other subjects as well as their own.

These results suggest that prior knowledge and awareness about the potential actions of one's partners (as is the case in the Experts group) increases the awareness of the self and also increases the need to monitor one's actions.

4. Discussion

4.1. Gaze and other body-movement functions: collaborative coordination and Performance Studies

Following the details in the Results section, the data indicates differences between what is described in studies on turn-taking in more traditional social interactions where speech is present and the results found within this particular social interaction of dance improvisation. The most surprising finding is the lack of mutual gaze, fundamental for joint-attention [29], which would be expected in a silent, collaborative decision-making context where turns need to be managed and coordinated.

Although a low number of communication-focused movements (i.e. body movements performed with representational intentions and having a symbolic or social nature) were initially predicted, given the participants' background and expertise in performance, it was not expected that there be absolutely none. Zero (n= 0) communication-focused movements, were found in the data. Data from a parallel study involving non-performers indicate a greater number of communication-focused MUs in this group [30]. One explanation for this is that non-performers will fall back on those non-verbal turn-taking strategies commonly used in conversation when they know they cannot use speech (mutual gaze, facial expressions, etc.), whereas these performers have embodied other strategies that non-performers do not have available.

One out of every three self-focused MUs were produced with fingers and one or both hands (n= 58; 33.5%) and are comparable to what are described in the literature as self-adaptors. Self-adaptors are body movements, such as scratching or fidgeting, typically produced under more "stressful" conditions because their production has a self-regulating and soothing function (inter alia [31, 32, 33]). Although these performers are experts in their domain, there is nonetheless a cognitive load as they must determine what, how, when, and if to improvise next. This may explain the high number of self-focused movement units across participants.

So how is it possible that these performers collaborated and coordinated without communication-focused gestures or even gaze exchanges, let alone made an improvised composition in real time? We posit that the greatest amount of information came via peripheral vision (e.g. [34]). This idea is based on our observations (such as participants' blinking patterns), but also from self-reporting from the choreographer Fiadeiro himself. In the context of this improvisation exercise, the types of visual inputs are quite limited, and to avoid "stealing the stage", performers monitor their body movements. With regards to turn-taking, using parafoveal and peripheral vision is sufficient to detect movements, such as if someone suddenly gets out of her chair to perform.

Covert attention [35] is most likely activated by the Expert performers as a strategy while they are fixating on the Game Table without ostensibly and unnecessarily moving their heads. In other words, the Expert performers intentionally allocate their attention following the goal or task they have at hand. This endogenous orienting [36] requires broadening the scope of perceptual attention, which in turn, may affect creativity by generating more original and extra-categorical uses for the Game objects in this improvisation [37].

Gaze is sometimes treated as if it were an autonomous behavior, where the eyes just move depending on what attracts them. We would like to highlight a fact which often goes unnoticed: gaze in fact is often controlled and monitored by the person according to their context. Thus, gaze becomes an im-

portant part of the social interaction context, and much like metaphors and gestures, it is a "structuring structure" which becomes part of the speaker-gesturer's conceptual system [38, 39]. We would like to suggest gaze's relationship to the concept of Bourdieu's "habitus" [40] and practice theory, a discussion which will be addressed in another venue. The data indicates that the performers did not use mutual gaze as a primary strategy in this collaborative process or in the turn-taking. It seems that performers use gaze as a habitus and an embodied practice, in the more sociological term. This avoidance of intersubjectivity in the traditional sense is paradoxical. It is not motivated only by the desire to be focused on the creative and collaborative task of the improvisation (cognitively akin to what is described in [41]). It also serves a social function of showing the others that they are participating in that very practice of performance, and that they are adhering to what is expected of an expert performer in that context. By avoiding intersubjectivity, they are being connected with their dance partners, who expect that type of behavior, hence forming a coordinated communicative behavior [42].

One of the more interesting findings for performance studies concerns shared attention and collaborative coordination during a creative sequence. The data displays few and brief moments where everyone's gazes converge onto the same focal point simultaneously, an indication that the individuals were commonly attending to different things. There is only one longer significant moment which lasts some twenty-seven seconds where all gazes meet on the Game table. Before this stretch of time, a number of improvised actions had already been enacted by various artists. When the last action was performed, there is almost no noticeable movement activity in any of the participants' bodies and all performers are attending to what has just happened for almost half a minute, a considerable amount of time. The other performers not only may be thinking of the future (what next action could be improvised), but they are also appreciating the present moment, as evidenced by a smile which emerges on the fourth performer's face.

Fiadeiro describes a phenomenon of "real time suspension" in his method, where dancers "accept that the creative flow is suspended and that they are together suspended in the flow" (personal communication). This may well be one of these moments, as everyone is looking at their joint creation, hardly moving, until one of the dancers decides to build on the creation. What seemingly is a moment of sacred silence in the creative process, with minimal movements and the group's fixated gaze, may well be an indication of collaborative coordination. Analyses like these may allow us to use group behavioral data to better identify moments of creativity and collaboration in other research.

4.2. Decision-making and precursory gestures

A phenomenon which emerged during the analysis of this decision-making exercise of the Game, and which is not entirely described in the literature (cf. [43]), is what we have dubbed the *precursory gesture*, in that it is a tell-tale movement of the gesture that is (or was) to come.

On various occasions, before participants were going to perform an action, they made a rapid and small body movement having the qualities of a preparation phase [44], followed by an immediate retraction or a hold. For example, before moving the hand and the arm to perform an action, there would be a small movement in the same hand, moving along the same path and direction, and with the same hand shape as the subsequent gestural movement. This type of body movement could be inferred as a hesitation in the decision-making process, including the decision of whether to self-select oneself in turn-taking.

We speculatively define the precursory gesture as a gesture, or more precisely a body movement, which is imagistically and functionally related to its more complete, immediate *successor gesture*. Here we tentatively describe its anatomy, function and timing, based on the observations from our data.

The anatomy of a precursory gesture is partial and not "well-formed" [45]. It is an incomplete and reduced image of the more complex successor gesture, sharing certain formal parameters. It typically includes a retraction phase characterized by the relaxation of the muscles involved in the gesture's production and a return to its initial rest position.

In terms of function, the successor gesture executes an intentional, or directed and purposeful, action; the precursor embodies the initial hesitation to perform that action.

As for the timing, our data indicates that the duration of the precursory gesture is speculatively and generally on the order of hundreds of milliseconds, and the successor occurs after a time on the order of seconds; however, these times are relative to the size of the articulator, by virtue of the physics of larger masses (for example in precursory gestures produced by the torso versus a hand).

As opposed to other types of communication-focused human gestures, which are referential and/or representative [46], we posit that the precursory gesture is not at all symbolic; rather, it is a self-focused, neuro-physiological bodily response to an uncompleted intentional action. These gestures may very well be universal if the function is tied to the biology of the gesturer and not to a symbolic system, and analogies are present in non-human primate data (Hélène Cochet, personal communication). Precursory gestures, which we will describe more in depth in future publications, were recurrent in our data and might prove useful in other research on multimodality and decision-making. Further research is recommended to better define and clarify this phenomenon.

4.3. A note on Cohen's kappa, contingency tables, and detecting gross errors

In Gesture research, observations of non-verbal behavior are typically conducted by a close inspection of video-recordings and displayed as spatio-temporal segments on a timeline in one of the available annotation tools. The segmentation and annotation work are conducted by independent human raters, who determine the beginning and the end of the gesture movements, as well as assign labels from an annotation scheme to the segments. Exactly this decision-making process of segmentation and annotation work creates problems in calculating the value of inter-rater agreement (IRA), and thus in estimating the validity and reliability of the collected material. Although various statistical coefficients are currently used in the measurement of IRA (e.g. Fleiss' kappa [47], Krippendorff's alpha [48]), Cohen's kappa [49] still remains the mostly widely used statistical measurement in the field, mainly because the kappa value informs the researchers on raters' agreement, disagreement and their agreement by chance.

To reach IRA, we calculated a modified Cohen's kappa using a function in the ELAN software. The determination of the inter-rater reliability in the tool is based on an algorithm by [18], which has the advantage of considering not only the raters' annotation agreement but also their segmentation agreement. The IRA output presents tabular results of cross-matched annota-

Table 3: Extract of the contingency table for gaze. The dark gray cell marks the gross errors committed because of "annotator fatigue", the gray cell displays correctly matched labels between two annotators. The diagonal in light gray highlights positive crossing between raters.

	gaze_to_home	gaze_to_gameT	gaze_to_objectT	gaze_to_p1	gaze_to_p2	gaze_to_p3	Unmatched
gaze_to_home	0	0	0	0	0	0	0
gaze_to_gameT	0	22	18	0	0	0	5
gaze_to_objectT	0	0	10	0	0	0	0
gaze_to_p1	0	0	0	2	0	0	0
gaze_to_p2	0	0	0	0	14	0	3
gaze_to_p3	0	0	0	0	2	4	1
Unmatched	1	4	5	0	3	2	0

tions between two annotators (contingency tables), as well as values of agreement by chance (modified Cohen's kappa), pure raw agreement, and the Kappa maximum (see Tab. 1).

The contingency tables were used in this study as a methodological tool in the annotation process to detect gross errors committed because of "annotator fatigue". The matrix table for gaze (Tab. 3) exemplifies how a gross error was made evident after a brief examination of the table. Because of this unusual value outside of the diagonal, both raters consulted the data and noticed that one of the raters had wrongly assigned the label "gaze to objects table" (dark gray cell) 18 times as "gaze to game table" (gray). Since the two tables were placed distant from one another, this mistake cannot be considered as an interpretational misjudgment and counts simply as a gross error, rectifiable without affecting the data. We advocate the use of this procedure to eliminate any similar mistakes resulting from annotator fatigue. In our case this supported us in reaching high IRA and in confirming the reliability of our data and subsequent analyses.

5. Conclusions

This study intends to contribute to the existing literature on turn-taking, presenting a novel context, that of a contemporary dance improvisation, which is multi-party and absent of any verbal communication. Unlike in more common, everyday social interactions, we found that intersubjectivity was actively avoided during this performance of the contemporary dance improvisation of João Fiadeiro's Real Time Composition Game, both in the performers' bodily movements and mutual gaze. We extrapolate that peripheral vision was chiefly deployed as a regulating strategy by these experts during the performance to coordinate turn-taking, but social practice and habitus also played a heavy role. The data provides zero cases of communication-focused movements. Although context- and communication-focused movements were monitored by the performers, self-focused movements seemed less monitored and were in fact overwhelmingly present, a further indication that these bodily movements are produced as neurophysiological responses to a cognitive load (self-adaptors).

In the qualitative analysis, we compared the data from the Expert performers with analogous data from Non-Performers introduced to the Game. A macro-analysis of the data frames the observations under the light of recent social cognition and decision-making literature.

Furthermore, we identified a class of body movements occurring in decision-making contexts that we have dubbed "precursory gestures", and we describe the anatomy, function and timing of these bodily movements.

From a methodological perspective, we argue for using the modified Cohen's kappa (notwithstanding its shortcomings) to validate researchers' annotation schemes and to achieve inter-rater agreement between two annotators. We also advocate using contingency tables as a tool to correct for "annotator fatigue" by highlighting gross errors.

6. Future research

The current paper reports preliminary results of the data collected in a contemporary dance improvisation. Future research will compare and contrast the data from the Experts group with a Non-Performers group, focusing on gaze and body movement units and their functions, in particular self-focused ones. These data will be further analyzed within the context of sociological practice. We also intend to investigate in greater detail the phenomenon of the precursory gesture in decision-making contexts, such as turn-taking.

A broader qualitative analysis of the collected data, with a special focus on participants' individual differences, is planned. We would like to closely examine how the body "reacts" and which non-verbal signals are observable across the groups at different stages of collaborative decision-making processes such as the one presented here.

Furthermore, we aim to detail a computational model tool for the visualization of eye gaze and MU data [50] in order to better evaluate annotated data.

7. Acknowledgments

This work was supported by the European Research Council under the project (Ref. 336200) - "BlackBox: A collaborative platform to document performance composition: from conceptual structures in the backstage to customizable visualizations in the front-end". The authors would like to thank Silvia Almas, Gina Joue, Conceição Amado, Mariana Escudeiro, Luís Correia, and the other members of the BlackBox team for their help and input. We are grateful to an anonymous reviewer for offering many helpful comments. A very special thanks goes to the choreographer João Fiadeiro, the RE.AL Atelier staff, and all the performers who participated in this study.

8. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, 1974.

[2] J. Fiadeiro, "Wenn du das nicht weißt, warum fragst du dann?" in *Wissen in Bewegung. Perspektiven der künstlerischen und wissenschaftlichen Forschung im Tanz*, G. Sabine, P. Husemann, and von Katharina Wilke, Eds. Bielefeld: transcript Verlag, 2007, pp. 103–112.

[3] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe, "Multimodal analysis of expressive gesture in music and dance performances," in *Gesture-based communication in human-computer interaction*. Berlin: Springer, 2004, pp. 20–39.

[4] E. Goffman, *Behavior in Public Places*. New York: Free Press, 1963.

[5] V. Petukhova and H. Bunt, "Who's next? Speaker-selection mechanisms in multiparty dialogue," in *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue*, Stockholm, 2009, pp. 19–26.

[6] D. Bohus and E. Horvitz, "Multiparty turn taking in situated dialog: Study, lessons, and directions," in *Proceedings of the SIGDIAL 2011 Conference*, Portland, Oregon, 2011, pp. 98–109.

[7] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[8] E. Ochs, E. A. Schegloff, and S. Thompson, "Interactional units in conversation: Syntactic intonational and pragmatic resources for the management of turns," in *Interaction and Grammar*. Cambridge: Cambridge University Press, 1996, pp. 134–184.

[9] L. Mondada, "Multimodal resources for turn-taking: pointing and emergence of possible next speakers," *Discourse Studies*, vol. 9, no. 2, pp. 194–225, 2007.

[10] L. Cerrato and M. Skhiri, "Analysis and measurement of head movements signalling feedback in face-to-face human dialogues," in *Proceedings of the First Nordic Symposium on Multimodal Communication*, P. Paggio, K. Jokinen, and J. Jönsson, Eds., Copenhagen, 2003, pp. 43–52.

[11] B. Rahayudi, R. Poppe, and D. Heylen, "Twente debate corpus - a multimodal corpus for head movement analysis," in *International Conference on Language Resources and Evaluation (LREC 2014)*, vol. 106, Reykjavik, 2014, pp. 4184–4188.

[12] J. B. Bavelas, "Appreciating face-to-face dialogue," in *Audio-Visual Speech Processing*, British Columbia, 2005, vol. 1.

[13] K. Jokinen, "Gaze and gesture activity in communication," in *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, C. Stephanidis, Ed. Berlin: Springer, 2009, pp. 537–546.

[14] G. Brône, K. Feyaerts, and B. Oben, "On the interplay between eye gaze and speech," in *Empirical Approaches to Multi-Modality and to Language Variation (AFLiCo 5)*, Lille, France, 2013.

[15] K. Shockley, D. C. Richardson, and R. Dale, "Conversation and coordinative strucutres," *Topics in Cognitive Science*, vol. 1, pp. 305–319, 2009.

[16] J. Holler and K. H. Kendrick, *Gesture, gaze, and the body in the organization of turn-taking for conversation: Insights from a corpus using new technologies*. Los Angeles: 4th International Conference on Conversation Analysis (ICCA14), 2014.

[17] C. S. Peirce, "Philosophical writings," in *Logic as Semiotic: the Theory of Signs*, J. Buchler, Ed., Dover, NY, 1955 [1902], pp. 98–119.

[18] H. Holle and R. Rein, "The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation," in *Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour*, H. Lausberg, Ed. Frankfurt am Main: Peter Lang Verlag, 2013, pp. 261–277.

[19] R. J. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[20] C. Oertel, M. Włodarczak, J. Edlund, P. Wagner, and J. Gustafson, "Gaze patterns in turn-taking," in *Proceedings of Interspeech 2012*, Portland, Oregon, 2012.

[21] C. D. Frith and T. Singer, "The role of social cognition in decision making," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 363, no. 1511, pp. 3875–3886, 2008.

[22] N. Sebanz, D. Rebbechi, G. Knoblich, W. Prinz, and C. D. Frith, "Is it really my turn? An event-related fMRI study of task sharing," *Social Neuroscience*, vol. 2, no. 2, pp. 81–95, 2007.

[23] R. Rico, M. Sánchez-Manzanares, F. Gil, and C. Gibson, "Team implicit coordination processes: A team knowledge-based approach," *The Academy of Management Review*, vol. 33, no. 1, pp. 163–184, 2008.

[24] R. Fusaroli and K. Tylén, "Carving language for social coordination: A dynamical approach," *Interaction Studies*, vol. 13, no. 1, pp. 103–124, 2012.

[25] B. Skyrms, Ed., *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press, 2004.

[26] R. A. Shweder, M. Mahapatra, and J. G. Miller, "Culture and moral development," in *Cultural psychology. Essays on Comparative Human Development*, J. W. Stigler, R. A. Schweder, and G. Herdt, Eds. New York: Cambridge University Press, 1990, pp. 130–204.

[27] Y. Tsushima, Y. Sasaki, and T. Watanabe, "Greater disruption due to failure of inhibitory control on an ambiguous distractor," *Science*, vol. 314, no. 5806, pp. 1786–1788, 2006.

[28] T. Singer and E. Fehr, "The neuroeconomics of mind reading and empathy," *American Economic Review*, vol. 95, no. 2, pp. 340–345, 2005.

[29] S. J. Rogers, "Mutual gaze," in *Encyclopedia of Autism Spectrum Disorders*, F. R. Volkmar, Ed. New York: Springer, 2013, pp. 1966–1967.

[30] V. Evola, J. Skubisz, and C. Fernandes, "The role of gaze and other body movements in collaborative decision-making: A study on coordinating turns in a contemporary dance improvisation exercise," in *7th Conference of the International Society for Gesture Studies (ISGS 2016)*, Paris.

[31] E. Paul and F. Wallace V, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[32] M. H. Krout, "Autistic gestures: An experimental study in symbolic movement." *Psychological Monographs*, vol. 46, no. 4, pp. i–126, 1935.

[33] H. Lausberg and H. Sloetjes, "Coding gestural behavior with the NEUROGES-ELAN system," *Behavior Research Methods, Instruments, & Computers*, vol. 3, no. 41, pp. 841–849, 2009.

[34] M. Gullberg and K. Holmqvist, "Keeping an eye on gestures. Visual perception of gestures in face-to-face communication," *Pragmatics & Cognition*, vol. 7, no. 1, pp. 35–63, 1999.

[35] M. Carrasco, C. Penpeci-Talgar, and M. Eckstein, "Spatial covert attention increases contrast sensitivity across the CSF: Support for signal enhancement," *Vision Research*, vol. 40, no. 10, pp. 1203–1215, 2000.

[36] M. I. Posner, "Orienting of attention," *Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[37] R. S. Friedman, A. Fishbach, J. Förster, and L. Werth, "Attentional priming effects on creativity," *Creativity Reserach Journal*, vol. 15, no. 2 & 3, pp. 277–286, 2003.

[38] V. Evola, "Multimodal cognitive semiotics of spiritual experiences: Beliefs and metaphors in words, gestures, and drawings," in *Form, Meaning, and Body*, F. Parrill, V. Tobin, and M. Turner, Eds. Stanford: CLSI, 2010, pp. 41–60.

[39] ——, "Metafore, sistemi religiosi e semiotica cognitiva multimodale. discorsi, gestualità e disegni di predicatori di strada cristiani e di un satanista," *Sistemi Intelligenti*, no. 1, pp. 23–48, 2010.

[40] P. Bourdieu, *Outline of a Theory of Practice*. Cambridge: Cambridge University Press, 1977.

[41] M. Kimmel, "Embodied (micro-)skills in tango improvisation: How a collaborative behavioral arc comes about," in *Out for a walk. Das Entgegenkommende Denken, Actus et Imago. Berliner Schriften für Bildaktforschung und Verkörperungsphilosophie*, F. Engel, S. Marienberg, and P. Schneider, Eds. Berlin: De Gruyter, 2015, vol. 15, pp. 57–74.

[42] L. Brandt, "Dance as dialogue: Metaphorical conceptualization and semantic domains," in *Sémiotique de la Musique - Music and Meaning (Signata 6)*, J. R. do Carmo Jr and P. A. Brandt, Eds. Liège: Presses Universitaires de Liège, 2015, pp. 231–249.

[43] H. Lausberg, "The NEUROGES coding system: Design and psychometric properties," in *Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour*, H. Lausberg, Ed. Frankfurt am Main: Peter Lang Verlag, 2013, pp. 85–106.

[44] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[45] ——, "Introduction," in *Language and Gesture: Window into Thought and Action*, D. McNeill, Ed. Cambridge: Cambridge University Press, 1992, pp. 1–10.

[46] I. Mittelberg and V. Evola, "Iconic and representational gestures," in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38.2.)*, C. Müller, A. Cienki, E. Fricke, S. L. Ladewig, D. McNeill, and J. Bressem, Eds. Berlin/Boston: De Gruyter Mouton, 2014, pp. 1732–1746.

[47] J. J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 5, no. 75, pp. 378–382, 1971.

[48] K. Krippendorff, *Content Analysis. An Introduction to Its Methodology*, 2nd ed. Thousand Oaks: Sage, 2004.

[49] J. Cohen, "A coefficient of agreement for nominal scales," *Education and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[50] C. Ribeiro, V. Evola, J. Skubisz, and R. Kuffner, "Transposing formal annotations of 2D video data into a 3D environment for Gesture research," in *7th Conference of the International Society for Gesture Studies (ISGS 2016)*, Paris.

An investigation of the effect of beat and iconic gestures on memory recall in L2 speakers

Eleni Ioanna Levantinou[1], *Costanza Navarretta*[2]

[1] Copenhagen University
[2] Copenhagen University
vpz728@alumni.ku.dk ,costanza@hum.ku.dk

Abstract

Previous studies have showed that iconic hand gestures aid memory recall and support comprehension in adults and children native speakers. In this paper, we investigate whether gestures might have an assisting role in second language acquisition. Repeating a previous experiment formed on native speakers, we used three types of stimuli (list of words accompanied by iconic gestures, beat gestures and no gestures) in order to test in which state the participants remember the most words. The result was that iconic gestures, compared to the other two states, provided significant support in memory recall and comprehension. However such an effect was not found with beat gestures whose presence gave worse results than the condition where no gestures were provided. This may indicate that beat gestures augmented the cognitive load of L2 speakers who have not learned yet how to interpret them.

1. Introduction

People use hand movements when they speak and these movements have many functions such as aiding comprehension or emphasizing the meaning of the discourse. This paper is about gestures and the meaning they convey along with speech and whether or not the use of gestures improves or ameliorates word recall. It appears that second language acquisition is a less focused on area and this raises the question of the validity of the assumption that gestures are an asset in any learning process. It is crucial to look at the role of gestures in second language acquisition and if they have an impact on the learning process and memorization. The main reason for this is that comprehension in a second language is made more difficult by several factors such as lack of vocabulary, different syntax, pronunciation and intonation differences. Therefore the inclusion of gestures in second language acquisition can possibly be very important.

Getting inspiration from So et al. [1], we will investigate the assistance that iconic and beat gestures may give in regards to L2 state. Similarly to So et al. [1] who investigated the assistance that gestures provide in native speakers' recalling ability, we replicate the experiments with non-native speakers of English.

The aim of this study is to investigate whether or not iconic and beat gestures aid memory and improve recall in L2 state, as they do for native speakers. Iconic and beat gestures have a different usage in language though. Iconic gestures are related to the meaning of the word with which they co-occur. On the other hand, beat gestures are not related to the meaning of the words with which they co-occur. They accompany the rhythm of speech and are aligned to the intonation of the language. Since intonation of a foreign language can be difficult to learn, when people speak a second language, they often use the intonation of their mother language. Indeed, when So et al. [1] investigated the impact of beat gestures in memory recall of children, they found that children were not assisted by the presence of beat gestures. Since children have not learned to interpret beat gestures which are related to information structure and the rhythm of the language, it is obvious that their memory will not get support from them. We assume that we will find the same tendency in second language speakers that is beat gestures will not aid comprehension and memory recall.

Therefore, we expect that the participants will remember more words represented with iconic gestures and less with beat gestures or words which are not accompanied by gestures at all. We anticipate that the percentage of the recalled words accompanied by iconic gestures will be higher than the percentage of the words recalled in the other two states.

This paper is divided into four main sections. The first part (section 2) concerns studies on gestures in general and on comprehension and memory specifically. In this part we also introduce the main hypothesis. The second part, section 3, concerns the design of the experiment in which we elaborate on the methodology of the study. The third part, section 4, consists of the results and findings of the research and the final part is the discussion (section 5). The discussion summarizes the results and compares them with other related works that may explain them. Furthermore, it proposes future work.

2. Related literature

Much research has been made in regards to gestures in general and to how they aid memory and recall. Ekman and Friesen [2] first categorized hand gestures in three types based on the "origin, coding and usage of the act": emblems, illustrators and adaptors. Building on these findings McNeill [3] introduced the categories of hand gestures: "iconic, metaphoric and deictic gestures". Iconic gestures represent the meaning with specific movements that are related to the meaning of the word and help convey it to the listener. Indeed Kendon [4], Alibali et al. [5] and Ozyurek [6] proved that iconic gestures serve a major communicative role and they aid listeners' comprehension. Metaphorical gestures convey meaning in a more abstract way. McNeill [3] specified some types of metaphoric gestures and he introduced "batonic (beat)" gestures. "*These movements are rhythmic and the movement is a simple up and down motion." (McNeill, [3], p. 84).* This distinction is relevant because it points out the difference between beat gestures and other non-representational gestures. Since many gestures are produced

unconsciously and speakers do not think before they produce them (McNeill [3]), it is important to see whether these gestures actually help the listener understand the message. Jacobs and Garnham [7] studied the effect of gestures in a narrative task. They found that gestures produced throughout the time of narration assist listener's comprehension. They proved that in a narrative task gestures play a major role in the overall understanding of the story. Alibali et al. [5], Kendon [4] and Ozyurek [6] studied the role of gestures in speech production and they concluded that the different kinds of gestures are different in execution but serve the same purpose: to convey the meaning efficiently and to assist the listener to get a better grasp of the meaning. Furthermore, Goldin-Meadow et al. [8] also showed that gestures may provide assistance in serving the meaning in a way that speech alone is unable to convey. Research into the communicative aspect of gestures begins from the point that important information can be conveyed non-verbally. Abundant research demonstrates the communicative role of gestures. Gestures constitute non-verbal cues that facilitate problem solving (Goldin-Meadow et al. [9], Kelly et al,[10], [11]) and they help disambiguation of similar terms which can be also referred to as "lexical discrimination" (Tompson and Massaro,[12]). One more relevant point in the communicative role of gestures, also mentioned before, is how enactment supports narrative processing (McNeill, Cassel and McCullough, [13]). Recently the mnemonic aspect of gestures has also been increasingly studied. Kelly, Barr, Church and Lynch [14] conducted experiments in which they investigated the impact that

In a broader approach, Tellier [24] investigated the role of gestures in teaching and she pointed out the importance of using gestures during teaching. In 2008 [25], she focused on second language acquisition and she showed that gestures assist reproduction of knowledge and help memorization in L2 state. She indicated that multimodality favors memory since gestures are not only a visual modality but also a motor one which have as a result richer trace in memory. Not only Tellier but also Gullberg along with others pointed out the importance of gestures in second language acquisition (Gullberg [26]; Gullberg et al. [27]; Macedonia et al., [28], Morett [29]).

Finally, So et al [1] studied the mnemonic effect of iconic and beat gestures in adults and children and they asked whether or not gestures need to be meaningful in order to facilitate recall and promote memory. They tested two groups of people, adults and children, on the same task adjusted to their abilities, in representational and non-representational gestures. They reached the conclusion that enactment enhances memory in both adults and children. Recall was better in the case of iconic and beat gestures for adults, but only in case of iconic gestures for children. They remembered more words accompanied with iconic gestures than when they were accompanied with beat gestures.

The evidence stated above suggests that at least to some extent it is proven that gestures aid memory and recall. Building on these previous works and especially So et al.'s [1] experiments, we would like to evaluate the tendency that subjects may have to show greater recalling abilities when exposed to specific kinds of gestures within the context of second language acquisition.

3. Methodology

gestures may have on comprehension and memory. They showed that speech and gestures work together and equally help to channel a meaning. Equivalently, Cook, Mitchell and Goldin-Meadow [15] found that when children gesture during the learning process of a new concept, gestures help preserve new information. In this study, members of the control group, which did not use gestures, were not so effective in memorizing the new task and this showed that gestures are a tool for increasing children's ability to memorize. Similarly, Stevanoni and Salmon [16] focused on new knowledge storage and found that gesturing promote and support learning and the process of memorizing. Furthermore, Goldin-Meadow and Wagner [17] supported that gestures provide a profound knowledge into the speakers' thoughts and they pointed out that gestures are an effective tool in learning, comprehension and memory. They gave evidence that they help speech to convey meaning and support memory. Dual Coding Theory (Clark & Paivio [18]) is a related theory which also supports the idea that multimodal learning reinforces memory. This has been also introduced before by Baddeley [19] who said that the role of gestures is to effect memory in a more efficient way. A lot of studies have been made on multimodality in learning and they have shown that multimedia learning is more efficient since two parts of the brain are involved: the auditory and the visual (Moreno & Mayer [20]). In addition, gestures enhance the trace in memory and make it stronger and more efficient and they assist the process of recall (Engelkamp and Cohen [21]; Cohen and Otterbein [22]; Nyberg [23]).

We replicated the methodology of So et al.'s [1] study on non-native speakers of English. In our experiment, we used three videos in each of which a native English speaker says a list of 10 words. In the first video the narrator uses iconic gestures, in the second beat gestures, and in the third, no gestures at all. The words are the same as those used in So et al.'s [1] study (see Table 1).

Table 1. *Lists of words.*

List 1	List 2	List 3
Come	Write	Cycle
Think	Strum	Run
Fly	Cut	Read
Comb	Walk	Carry
Stir	Swim	Inject
Pray	Throw	Climb
Stack	Turn	Bounce
Beg	Eat	Brush
Hammer	Open	Knock
Cry	Push	Listen

All words are English verbs of one or two syllables. So et al. use three lists because they test the three conditions on the same participants and memory would be affected if the same list is repeated three times. We have reused their list in order to be able to compare our results with theirs. The duration of the videos is the following: 33sec the iconic gestures video, 26sec the beat gesture video and 22sec the non-gesture video. Both iconic and beat gestures last approximately 3msec.

Two groups of participants were tested. They were all university students, aged 24-35. The first group was the

control group and was composed of 4 native English speakers. They were tested on the same task in order to test whether we could replicate the results and justify the findings in So et al. [1]. If we can obtain similar results the control group as those obtained by So et al. then it would be reasonable to perform the same test on English non-native speakers.

The second group was composed of 10 non-native speakers of English (4 males and 6 females), and was used to test the hypothesis. The non-native speakers, have a high level of English as they are all currently enrolled in an international master at the university of Copenhagen where English is the teaching language.

Both groups had to follow the same procedure. The participants were asked to see three videos. In each video the English narrator went through the list of English verbs. In the first video the narrator accompanied the words with iconic gestures (i.e. gestures which visually represent the meaning of the verbs in real time), in the second video, he said the words accompanying them with beat gestures[1] and in the third video he said the words without performing any gesture. Iconic gestures were chosen based on how often they were used by a native speaker to accompany these words.

After each video the participants were asked to recall as many words as they could without any time limitation. Furthermore, the participants were asked to hold a pen in order to inhibit gesturing both while they watched the video and while they recalled the words. The reason for this is that if they gestured during the playback of the video, gesturing could have helped word memorization and if they gestured during recall, gesturing could have facilitated retrieval of the words according to the Lexical Retrieval Hypothesis (Holler et al. [30]).

Moreover, participants were not allowed to repeat the words during playback of the video. Between each video, the participants were asked to solve a simple mathematical task in order to prevent interference of the words between the conditions (So et al. [1]). By using an unrelated mathematical task, we also wanted to distract the participants from a linguistic retrieval mental process.

4. Results

The main hypothesis we wanted to test is that the participants would remember more words accompanied with iconic gestures, less with beat gestures, and the fewest with no gestures at all.

The analysis that we performed on the data was mainly made with SPSS along with some coded calculations via Python programs written for this purpose.

First, we counted the number of the correctly recalled words and we calculated the percentage of the recalled words for each condition. As shown in Table 2 the percentage of the words recalled from the iconic video is higher than the percentage of those recalled in the two other conditions. For the control group it is 73.3% and it is slightly higher than for the treatment group for which is 71%. However, the results obtained with 2L speakers indicate the same effect. In both groups, the percentage of the words recalled from the iconic gestures video is higher than the words recalled from the other two videos. For native speakers, the percentage of words recalled from the video with beat gestures is larger than the percentage of words recalled from the non-gesture video, which is opposed to the results for non-native speakers. These percentages are 50% for the beat gestures video and 40% for the non-gestures video for native speakers, while the respective numbers for non-native speakers are 37% and 48%. Thus, we obtained for our control group results similar to those obtained in the experiment by So et al. [1] that is beat gestures have a positive effect on memory recall for native speakers of English, but this effect is not as large as that provided by iconic gestures.

A first analysis of the data indicates that iconic gestures aid memory in English even when the participants are non-native speakers, but we do not have the same indication for beat gestures. On the contrary, the second language speakers recalled more words from the non-gesture video than from the beat gesture video.

Table 2. *Percentage of the words recalled*

	Iconic gestures	Beat gestures	Non gestures
Control group	73.3%	50%	40%
Test group	71%	37%	48%

In a second analysis, we calculated the minimum, the maximum and the mean of the words recalled. As it is shown in Table 3, the sample of the words is 10. The minimum number of recalled words accompanied with iconic gestures is 5, while with beat gestures it is 0 and with no gestures 3. On the other hand, the maximum number of recalled words accompanied by iconic gestures is 10, with beats and no gestures it is 7. The mean of the recalled words accompanied with iconic gestures is 7.1, with beat gestures the mean is 3.7 and for the non-gestures condition it is 4.8. Obviously, taking into consideration both percentage and mean, the figure suggest the same tendency as that provided by Table 2. Iconic gestures are more helpful for memorization and recall than beat gestures and no gestures. In the case of beat gestures and no gestures, recall is better when no gesture accompanies the words to be remembered.

Table 3. *Descriptive statistics*

	N	Min	Max	Mean	Std D
Iconic	10	5	10	7.10	1.449
Beat g	10	0	7	3.70	1.829
Non g	10	3	7	4.80	1.398

Subsequently, we used SPSS in order to investigate if there is a significant difference between the results obtained with the three different conditions. After checking our data for normality using Kolmogorov-Smirnov Test we applied non-parametric pair tests in order to determine the significant difference between the results for the three conditions.

In order to test the difference between the three conditions, we performed one way- Anova analysis. In this case the subjects of a group are measured in multiple comparisons under three different gesture conditions. The analysis has been made to compare the group's answers under the three different

[1] "*Beat gestures can take various forms of hand shapes and movements. One of the most typical forms of beat gesture is a hand with open palm flips outwards*" (McNeill, [3] referenced in So et al., [1], p.5)

conditions. The result will be whether or not significant difference (*Sig. value*) exists between the three conditions of the same group. If *Sig. value* is equal to or less than *.05* (e.g. . 035, .02, .005), then a significant difference exists between the three conditions of the same factor. (Pallant J. [31]).

Table 4. *Tests of Normality*

	Statistics	df	Sig.
Iconic g	.176	10	.200
Beat g	.265	10	.045
Non g	.316	10	.005

For the first two pairs, the iconic-beat gestures pair and the iconic-non gestures pair, the difference is statistically significant. For the last pair, beat-non gestures, the difference is not statistically significant.

Table 5. *P-values between the conditions*

	Iconic gestures	Beat gestures	Non gestures
Iconic g		.00001	.002
Beat g	.00001		.437
No g	.002	.437	

More specifically, the difference between the first pair, iconic-beat gesture, is statistically significant since *p= 0.0001*. The same goes for the second pair, iconic-non gesture. In this pair, *p=0.002* and it is clear that there is also a statistically significant difference here. The same cannot be said though, for the third pair, beat–no gesture. The difference between them is *p=0.437*.

The results indicate the same tendency among native and non-native speakers in regards to whether or not words accompanied with iconic hand gestures aid memory. The same tendency was not found among the groups as it concerns words accompanied with beat gestures or words alone.

Furthermore, our results demonstrate that words accompanied by iconic gestures are easier to be recalled by a non- native speaker than words that are accompanied by beat gestures or that are presented alone. We did not find a significant difference between the results obtained with beat gestures and no gestures and this suggests that in non-native speakers there is no difference as to whether we use beat gestures or not. However, the data indicate a tendency for having worse results in the beat condition than in the no gesture condition for 2L speakers.

5. Discussion

The purpose of the study was to investigate the impact of two different types of gestures on memorization and comprehension in second language speakers. The assumption that iconic gestures facilitate comprehension and recall was fully confirmed. Iconic gestures assisted memory recall in second language English speakers, as they did for native English speakers. Results showed that iconic gestures provide the same support to non-native speakers as they do to native speakers. Similarly to So et al's [1] findings on the impact that iconic gestures have on comprehension and memory, participants recalled more words from the first list, which were accompanied with iconic gestures, than from the other two lists of words which were accompanied with beat gestures or not accompanied at all. The explanation may lie in the fact that iconic gestures are representational of the meaning of a word (McNeill [3], Kendon [32]).

On the other hand, beat gestures or the absence of gestures, did not assist memory recall like the iconic gestures. Findings indicate a significant difference between the words recalled when they were accompanied by iconic gestures and beat gestures or words accompanied by iconic gestures and no gestures at all. Furthermore, comparing the number of words recalled accompanied by beat gestures and the number of the words recalled without any representation, we can see that participants remembered more words without any representation than with beat gestures. In explanation, participants claimed that beat gestures confused them and they got distracted from the words. For them hearing the words alone without any accompanying gesture was easier, and they were able to stay focused on the task. Thus it was easier to recall the words afterwards.

The difference of results obtained in the iconic gestures condition and the beat gestures condition is probably due to the fact that beat gestures are not representational, thus they do not convey or define the meaning of a word. As we have explained before, beat gestures co-occur with speech and are aligned to its rhythm. They follow the prosody of the language and contribute to information structure, i.e. they contribute to indicate new or important information. Since different languages have different patterns of intonation, it is often difficult for non-native speakers to follow the rhythm of the new language and discharge the rhythm of their native language. As a result, beat gestures aligned to a different intonation may cause cognitive overload and as a result they can be confusing for second language speakers. It is difficult for a non-native speaker to coordinate speech accompanied with beat gestures in a different intonation than the one they know and they are already used to. As our findings showed, words that had no distraction from the beat gestures were easier to remember (Gussenhoven [33]).

Nevertheless, our study also pointed out that iconic gestures have a positive impact in second language acquisition. Quinn and Allen [34] as well as Kelly et al. [35] have also demonstrated the fact that enactment assists learning of a second language. In fact, many researchers have shown that when iconic gestures co-occur with speech, they promote comprehension, improve memorization and assist memory recall (Tellier [36], [25]; Wagner, Nusbaum, Goldin-Meadow [37]). This happens because representational gestures make stronger impressions on the brain as two areas of it are involved, the auditory and the visual, thus the trace is more strong(Cohen & Bean, [38]; Clark & Paivio, [18]; Nilson & Craik, [39]).

The present study though, did not prove any positive impact on memory recall from beat gestures. Opposed to the findings of So et al. [1], beat gestures did not provide any assistance to adult non-native speakers of English. This needs further investigation and the new study should probably focus on beat gestures, but within the context of larger discourse context. If one or two words are emphasized in a sentence by beat gestures, they will probably be beneficial for memory and recall also for non-native speakers. The same task was used by Feyereisen [40] when he investigated the mnemonic effect of iconic and beat gestures within the context of a sentence. He

concluded that meaningful (that is representational) gestures benefit memory recall, more than non-meamingful (in the sense of non-representational) gestures. He also proved that non-meaningful gestures are more beneficial than no gestures at all. However, also Feyereisen [40] as So et al. [1] conducted the study on native speakers; therefore one suggestion is to perform an investigation of the same task on a second language. A suggestion for further investigation in the field of beat gestures could be a study with non-native speakers that have lived many years in the second country. These participants will be used to the intonation of the second country's spoken language and it would be interesting to investigate if beat gestures will provide assistance in memory and understanding.

In any case, the fact that gestures aid comprehension as they make listeners encode new knowledge to a more permanent and stable format has been demonstrated here, and in previous studies (Goldin-Meadow, Nusbaum, Kelly & Wagner [41]; Wagner, Nusdaum, Goldin-Meadow [37]) This was also proven in the case of learning (Bruken, Steinbacher, Plass & Leutner [42]; Mayer & Moreno [43]).

Further investigation could be also initiated in deictic gestures for second language comprehension: whether or not the relation of speech with the environment and pointing gestures towards objects will be helpful during the process of learning a new language (investigated in native speakers by Ballard, Hayhoe, Pook, & Rao [44]; Grant & Spivey [45]).

Furthermore, the present study focused on short term memory of non-native speaker adults, a suggestion for further research is to investigate long term memory in L2 or maybe to examine whether or not the same tendency can be found in young children. Additionally, since a language does not consist only of verbs, it would be interesting to explore the impact of gestures with other classes of words, like nouns and adjectives.

Since the current study was a pilot study, the nationality of each non-native participant was not taken into consideration, or how similar their mother language is with the English language, which was the language they tested on. Additional limitations are that the level of proficiency in English was not taken into consideration as well as the standard of pronunciation of the participants. Studying this may enlighten the obscure field of beat gestures since pronunciation is directly connected to the prosody and the intonation of a language and some languages have more similar prosody than others.

In conclusion, encoding words with gestures can be beneficial in second language acquisition since enactment aids comprehension and enhances memory recall. According to the findings, iconic gestures assist memorization but the same was not proven for beat gestures. Further research should be made in addition by looking at the impact of different kinds of gestures on second language acquisition.

References

[1] So W. C., Chen-Hui C. S., Wei-Shan J. L. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: is meaning in gesture important for memory recall? *Lang. Cogn. Process.* 27, 665–681

[2] Ekman, P. & Friesen W. V. (1972). Hand Movements. Journal of Communication*, 22*, 353-374.

[3] McNeill, D. (1992). Images, inside and out. In Hand and mind: What gestures reveal about thought (pp. 1135). *Chicago, IL: University of Chicago Press.*

[4] Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, 27, 175–200.

[5] Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. *Journal of Memory and Language,* 44, 169–188.

[6] Ozyurek, A. (2002). Do speakers design their co speech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language*, 46, 688–704

[7] Jacobs, N., & Garnham, A. (2006) The role of conversational hand gestures in a narrative task. *Journal of Memory and Language,* 26, 291-303.

[8] Goldin-Meadows, S., & McNeill, D. (1999). The role of gesture and mimetic representation in making language the province of speech. *In M. Corballis & S. Lea (Eds.)*, The descent of mind (pp. 155–172). Oxford: Oxford University Press.

[9] Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, 9, 201– 219.

[10] Kelly, S. D., & Church, R. B. (1997). Can children detect conceptual information conveyed through other children's nonverbal behaviors? *Cognition and Instruction,* 15, 107134.

[11] Kelly SD, Church RB. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development*. 1998;69:85–93.

[12] Thompson, L. A., & Massaro, D. W. (1994). Children's integration of speech and pointing gestures in comprehension. *Journal of Experimental Child Psychology,* 57, 327–354.

[13] McNeil, D., Cassell, J., & McCullough, K.E. (1994). Communicative effects of speech-mismatched gestures. *Language and Social Interaction,* 27, 223-237.

[14] Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language,* 40, 577592.

[15] Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, 106(2), 10471058.

[16] Stevanoni E., Salmon K. (2005). Giving memory a hand: instructing children to gesture enhances their event recall. *J. Nonverbal Behav.* 29, 217–233

[17] Goldin-Meadow, S. & Wagner, S. M. (2005). How our hands help us learn. Trends in Cognitive Sciences, 2005, *9*, 234-241.

[18] Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3(3), 149210.

[19] Baddeley, Alan (1990). Human memory: Theory and practice. *East Sussex: Lawrence Erlbaum Associates Ltd*. Publishers.

[20] Moreno, Roxanna & Richard E. Mayer (2000). A Learner-Centered Approach to Multimedia Explanations: Deriving Instructional Design Principles from Cognitive Theory. *Interactive Multimedia Electronic Journal of Computer Enhanced Learning,* 2 (2).

[21] Engelkamp, Johannes & Ronald L. Cohen (1991). Current issues in memory of action events. *Psychological Research*, 53, 175-182.

[22] Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Iconic and non iconic gestures compared. *European Journal of Cognitive Psychology*, 4(2), 113139.

[23] Nyberg, L. (2002). Levels of processing: A view from functional brain imaging. Memory, 10(5), 345-348. Retrieved October 26, 2009, from the *PsycINFO database.*

[24] Tellier, Marion (2006). L'impact du geste pédagogique sur l'enseignement-apprentissage des étrangères: Etude sur des enfants de 5 ans. *Unpublished Doctoral Dissertation.* University Paris 7 – Denis Diderot, Paris.

[25] Tellier, M. (2008). The effect of gestures on second language memorization by young children. *Gesture*, 8(2), 219235.

[26] Gullberg, Marianne (2008). Gestures and second language acquisition. In Nick C. Ellis & Peter Robinson (Eds.), *Handbook of cognitive linguistics and second language acquisition* pp. 276-305). London: Routledge.

[27] Gullberg, M., Roberts, L., Dimroth, C., Veroude, K., & Indefrey, P. (2010). Adult language learning after minimal exposure to an unknown natural language. *Language Learning*, 60, 5–24.

[28] Macedonia M,, Bergmann K., Roithmayr F.(2014) Imitation of a Pedagogical Agent's Gestures Enhances Memory for Words in Second Language. *Science Journal of Education*. Vol. 2, No. 5, 2014, pp. 162-169.

[29] Morett, L. (2014) When Hands Speak Louder Than Words: The Role of Gesture in the Communication, Encoding, and Recall of Words in a Novel Second Language. *The Modern Language Journal* Volume 98, Issue 3, pages 834–853, Fall 2014

[30] Holler J., Turner K. & Varcianna T. (2013). It's on the tip of my fingers: Co-speech gestures during lexical retrieval in different social contexts. *Language and Cognitive Processes*, 28:10, 1509-1518.

[31] Pallant Julie, (2007). SPSS Survival Manual. A Step by Step Guide to Data Analysis using SPSS for Windows. Third edition. *Open University Press*

[32] Kendon, A. (2004). Gesture: Visible action as utterance. *Cambridge University Press.*

[33] Gussenhoven, C., (2002). Intonation and interpretation: Phonetics and phonology. *In Proceedings of Speech Prosody* 2002, Aix-en-Provence.

[34] Quinn-Allen, L. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal,* 79, 521529.

[35] Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24(2), 313334.

[36] Tellier, M. (2005). L'utilisation des gestes en classe de langue: comment e´valuer leur effet sur la memorisation du lexique? In M. Billieres, P. Gaillard, & N. Spanghero-Gaillard (Eds.) *Actes du Colloque International de Didactique Cognitive, DidCog 2005*. Proceedings on CD-Rom, Toulouse.

[37] Wagner, S., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language*, 50, 395407.

[38] Cohen, R. L., & Bean, G. (1983). Memory in educable mentally retarded adults: Deficits in subject or experimenter. *Intelligence*, 7, 287298.

[39] Nilsoon, L. G., & Craik, F. I. M. (1990). Addictive and interactive effects in memory for subject performed tasks. *European Journal of Cognitive Psychology*, 2, 305324.

[40] Feyereisen, P. (2006). Further investigations on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, 18, 185Á205.

[41] Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science,* 12(6), 516–522.

[42] Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, 49, 109–119.

[43] Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual-processing systems in working memory. *Journal of Educational Psychology*, 90, 312-320

[44] Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4), 723–767.

[45] Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science,* 14 (5), 462-466

Top-down-Bottom-up Experiments on Detecting Co-speech Gesturing in Conversations

Kristiina Jokinen
Institute of Computer Science
University of Tartu
Kristiina.jokinen@ut.ee

Abstract

Automatic analysis of conversational videos and detection of gesturing and body movement of the partners is one of the areas where technology development has been rapid. This paper deals with the application of video techniques to human communication studies, and focuses on detecting communicative gesturing in conversational videos. The paper sets to investigate the *top-down-bottom-up* methodology, which aims to combine the two approaches used in interaction studies: the human annotation of the data and the automatic analysis of the data.

1 Introduction

Conversations form a social system whereby the interlocutors exchange information about their intentions, interests, and feelings. The participants use verbal and non-verbal means to give feedback and construct common understanding with their partner. Non-verbal communication (Argyle 1988) has been long studied focusing especially on gesturing (Kendon 2004), gaze (Argyle and Cook 1976), and various prosodic and paralinguistic issues (Schuller and Batliner 2013). However, it is only recently that advanced technology has given an opportunity to automatically detect these signals in such a robust way that also interaction studies can benefit of the objective views and of the automatic detection of signals; we talk about Social Signal Processing, which refers to data-directed statistical and machine-learning studies of the verbal and non-verbal signals exchanged in communication. Social signals indicate interest, emotions, affect, etc. and include a wide range of various behavioural signals like gesturing, gaze, laughing, coughing etc.

However, social signal processing requires large data-sets for enabling machine-learning studies and usually also golden standard corpora, or annotated corpora which provide reference point for the evaluation of the algorithms and models. Given the huge work and resource requirements for manual annotation, various algorithms and tools have been developed to assist in the initial analysis of the data, or conduct the segmentation automatically.

In this context, our studies also deploy novel technology in human communication studies, and explore the *top-down-bottom-up* methodological approach and its use in social signal processing. The aim is to provide an objective basis for human annotations concerning conversational partners' head, hand, and body movements, while also taking into account the interpretation of the events in their conversational space.

In particular, the paper focusses on video analysis and gesture recognition technology that enables observations of the human speakers and their movement on video recordings. The recognition technology, described in more detail in Vels and Jokinen (2015), decomposes the observed movement into three gesturing parts (body, head and feet), and regards them as separate activities. In this paper, this technology is used in human interaction studies, and the recognized gesturing is visualized together with the participants' speech, so as to correlate conversational participants' movements with their speaking and listening activity.

The paper is structured as follows. After a short introduction to the methodology in Section 2 and gesturing in Section 3, the paper presents the video processing technology and the data used in the experiments in Section 4. Results are discussed in Section 5, and conclusions and future work in Section 6.

2 *Top-down and bottom-up* Methodology

2.1 Top-down-bottom-up analyses

As already discussed in Jokinen & Pelachaud (2013), the *top-down-bottom-up* methodology for data annotation looks at communicative situation from two opposite viewpoints: the top-down approach is based on human observation and uses video recordings and manual tagging of the communicatively important events in the videos, according to some annotation scheme, while the bottom-up approach uses automatic technological means to recognize, cluster, and interpret the signals that the communicating agents emit.

Annotations also need to be consistent, and have the particular semantics they have been designed for, so the annotation results have to be validated by applying the scheme to practical coding tasks and by calculating inter-coder agreement by several coders (including also automatic coding algorithms). By combining the top-down approach, i.e. manual annotation and analysis of the data with the bottom-up analysis of the multimodal signals, it is possible to contribute to the validation of the data and to the quality of the annotated data given the data model and the annotation scheme. On one hand, automatic analysis lends itself to a basis for event detection, and on the other hand, manual annotation is used as a "gold standard" for clustering and classification tasks, to give semantics to the automatically found patterns.

To facilitate manual annotation, semi-manual annotation can be enabled by deploying supervised or unsupervised techniques as a preprocessing step. For instance, speech recognizers can be used to segment speech, parsers to provide linguistic knowledge, eye-trackers to trace gaze paths, and motion trackers as well as various face, gesture and body detection techniques to detect body movement and gesturing. The recognized events can then serve as candidates for more detailed communicative analysis, and the automatic techniques can thus assist human analysis, by segmenting the audio-video data in a uniform manner. Thresholds and parameter values must be set by experimentation and human judgement, but the systematic calculations can be said to produce an objective basis for further analysis, which helps to direct the initial segmentation on same level observations across the annotators, theoretical frameworks, activity types, and conversational settings. Although automatic recognition technologies require a *data model*, i.e. theoretical assumptions that describe the categories and classifications to be found in the data, it is, in principle, easier to determine required granularity and completeness levels by some measurable technical criteria than by more subjective conceptual definitions.

2.2 Internal intention vs external observation

An issue that needs discussion in this context is the very notion of the communicative meaning assumed to be carried by various social signals. It is possible to classify multimodal signals either by interpreting them as originating from the internal communicative intention of the participant, i.e. being displayed or signalled following Allwood's (2001) terminology, or by judging if the events have a noticeable effect on the recipient, i.e. based on the external annotator's observations on what happens in the situation. These two view-points result in different annotations since the former aims to model the participants' internal cognitive decisions, while the latter is based on the results of these actions. A similar distinction can be found in Speech Act theory (Austin 1962), where the notions of illocutionary and perlocutionary acts are introduced. Analogously in multimodal annotation, it is also possible to talk about two different types of annotations, depending on whether the analysis focuses on the *agent's internal intentions*, or on the *consequential effects of the agent's actions* upon the hearer.

2.3 Overt vs Covert meaning

However, even if the annotator is expected to select events that have a communicative function (either by looking at the item's illocutionary or perlocutionary force in the context), there is still another issue that needs attention, namely to determine if the item has an *overt or a covert communicative meaning*. It is well-known that spoken utterances can function as either direct or indirect speech acts, the latter referring to utterances that need deeper contextual inferencing to be correctly interpreted (cf. the classic example of requesting the opening of a window by stating that it is hot), and in a similar manner, multimodal signals can also be regarded as having direct or indirect meanings. For instance, emblems carry a direct, culturally specified meaning (which can be said to be indirect for those outside the culturally specified community), while pointing and iconic gestures directly identify and describe a referent. On the other hand, manipulative gestures, such as

lifting a coffee cup, rolling a pen, changing legs in standing position, etc. do not have an overt communicative function, yet they can indirectly demonstrate the agent's emotional state or intentional stance. They can be appropriately interpreted by the partner only if the partner has learnt to attend to such signalling and is able to draw appropriate conversational inferences to uncover the indirect meanings in the partner's gesturing. To reach the appropriate communicative inferences, the interlocutors need to understand the conversational situation and the principles that guide communication, i.e. they should distinguish the different level of conscious and intentional communication.

2.3 Intentions and segmentation

In human communication, often the difference between unintentional indication (e.g. blushing), intentional display (such as emphasising one's dialect when speaking), and conscious signalling (see Allwood 2001 for terminology) is difficult to determine, since it is difficult to determine the level of consciousness and volition that are behind the communicator's actions in the first place. In general, while it is possible to observe the partner's behaviour and make inferences on the possible reason and motivation for the various actions that the observer considers important in the given communicative situation, it may not be possible to fully understand, nor observe signals and actions in detailed enough manner to actually be able to understand, the actual reasons behind the partner's behaviour.

Human segmentations can thus differ widely depending on also what counts as a relevant event, and behaviour annotations can have different interpretations depending on what aspect of the action the annotator focussed on. The bottom-up approach, or pure signal detection without any particular linguistic knowledge about the meaning of the possible events, may come to help here. While signal analysis can provide rather detailed observations, it can also delay interpretation based on the level of granularity of the data analysis. The relation between form and function need not be one-to-one nor one-to-many, but many-to-many depending on the level of granularity chosen for the analysis in a particular context. The relevance of the various events may become clear only when the data patterns and clusters have been formed, and this can vary depending on the interpretation of the signals.

3 Gesticulation and gesturing

Following Allwood (2001) and Jokinen (2009), we consider interaction as a communication cycle where basic enablements of contact, perception, and understanding must be fulfilled in order for a full communication to take place. Often the enablements are signalled via multimodal signals, which thus form an integral part of the successful communication.

Kendon (2004) uses the term "gesture" to refer to visible action that participants distinguish and treat as governed by openly acknowledged communicative intent. The term "gesticulation" refers to the gesturing that occurs in association with speech and which is bound up with it as part of the total utterance. It consists of three phases (preparatory, peak, and recovery phases) that describe the different parts of the movement.

Interactive gestures form a class with the common function of including the listener in the conversation. They occur at specific moments in time and particular points in space, and can efficiently exert coordination of the conversation and provide meanings as the dialogue goes on.

Gestural signs are formed by the cognitive system that is also used in the movement of the body in the physical environment. Gesturing requires spatio-motoric thinking and ability to orient body parts toward a target in the physical environment, as well as the ability to track the target when it moves (Kita 2000).

Human body movements can be said to form a continuum from movements without any overt communicative meaning to movements which are communicatively significant gesturing. Body movements and the flow of speech are closely linked in human communication system and between the interlocutors. For instance, it is noted that the peak of the gesture coincides with the conceptual focal point of the speech unit, and each new representational gesture appears with a new unit of meaning. Both utterance and gesture derive from a deeper idea unit source that they represent co-expressively.

4 Data and recognition algorithm

For the experimentation we used the 23 dialogues from the MINT (Multimodal INTeraction) project collected at University of Tartu (Jokinen and Tenjes 2012). The speakers are unfamiliar with each other and make acquaintance with their partner for the first time (cf. Paggio et al. 2010).

Each file is about 5 minutes long and records the first encounter between the participants. There are 23 different participant (11 female, 12 male), and each person has dialogues with two different partners, i.e. appears in two videos. The partners face each other, and there are three cameras: one from front and two from sideways recording more on the partner's face from the front. Original Full HD (1920x1080 pixel) videos were resized to 640x360 px and 25 frames per second. A screen shot is given in Figure 1.

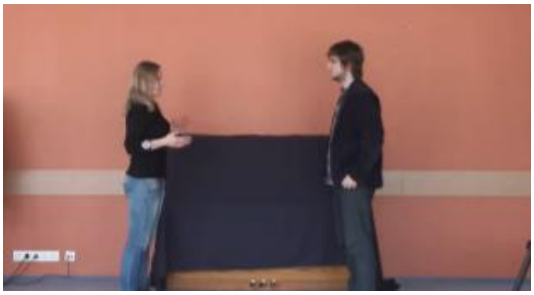


Figure 1 Screen shot from the MINT corpus.

Visual gesture movement recognition consists of several steps (Gonzales and Woods, 2010). On a general level these include:

a) Object representation: compare and decide on the suitable representation for the object tracking. An object can be represented e.g. by its shape or appearance.

b) Feature selection: choose visual features for tracking (colour, texture etc.)

c) Object detection: detect the object based on the chosen features.

d) Object tracking: log the movements of an object by tracking the trajectory of the set of features identified as the object.

e) Object recognition: interpret movements based on the analysis of gathered tracking information.

The previous algorithm (Vels and Jokinen, 2015a) allows us to find the positions of the moving persons in a video frame using a contour detection algorithm. It presented a novel idea of initializing a background model from a single frame using 8-neighbourhood of each of the frame pixel and randomly choosing 20 neighbour-pixels instances to build the model. In the follow-up paper (Vels and Jokinen, 2015b) the contour for the whole body is decomposed into head, torso, and legs bounding boxes so as to allow a more detailed analysis of the movements of the different body parts, by retrieving the precise coordinates of the bounding boxes which can be used to identify hand, head and lower body (foot) movements. Movements are also matched with speech events, which allows correlations to be analysed in easier and improves visualization of the conversational video.

Figure 2 shows a few screen shots of the results of the object segmentation process: background subtraction, morphological closing, body contour, and the final result. The colour video is converted first from RGB to grayscale, the Canny algorithm (Canny, 1986) is used for edge detection, and background subtraction is applied to recognize the objects from the background while morphological closing corrects border areas for final contours.

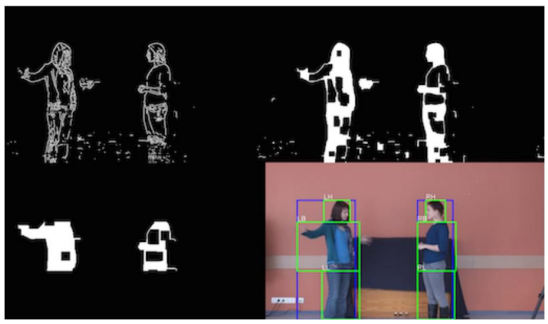


Figure 2 Four of the segmentation steps: background subtraction, closing, body contour, and final result with the detected head, body and leg coordinates.

Decomposition into head, hands, and foot bounding boxes starts by providing a very precise location and size of the head position, and then, using the relative position of the head with respect to the whole body, the body contour is located within which the coordinates for the torso and legs can be retrieved. Hand movement detection uses coordinates with noise removed. Median values of front and back coordinates of body surrounding boxes are used, and all values below a certain threshold are discarded. For the head coordinates, only the middle point value is used, as the head does not change its horizontal size. A simple peak detection algorithm is applied to the coordinates so as to retrieve possible hand movements.

The coordinates are recorded as follows: LBB (left person body back), LBF (left person body front), LH (left person head), RBB (right person body back), RBF (right person body front), and RH (right person head). With these coordinates we can capture all horizontal movements of the human head and body during the conversation.

5 Results and discussion

5.1 Speech and gesturing

The speaker's utterances were annotated by ANVIL manually for three categories: speech, laughs, and non-verbal vocalisations (e.g. *hmm, ahem*). The data is in XML-format and can be parsed automatically for calculating correlation with movement events, and for data visualization.

Figure 3 visualises synchrony of speech with body movements for about a half a minute long clip for the two speakers: the right speaker's movements are shown above, and those of the left speaker below. The right speaker (green coloured above) makes several rapid hand movements (lower green curve) during speaking (light green bars) with two non-verbal vocalisations (topmost dark green bars), but also seems to be rocking his whole body back and forth rhythmically (the green curves move simultaneously and in synchrony with the speech). On the other hand, the left speaker is rather still, and only one significant hand movement (upper blue curve) appears during own speaking (light blue bars). However, the left speaker produces non-verbal vocalisation (dark blue bars) and laughs (top-most dark blue bars) regularly, interleaving them with the partner's speech, and suggesting that the left speaker listens to the partner's lively spoken presentation and gives a lot of feedback to this. This exemplifies cooperation and synchrony between the speakers, and nicely confirms the hypothesis that the speaker moves more than the listener, and that the movements are synchronised (Battersby 2011).

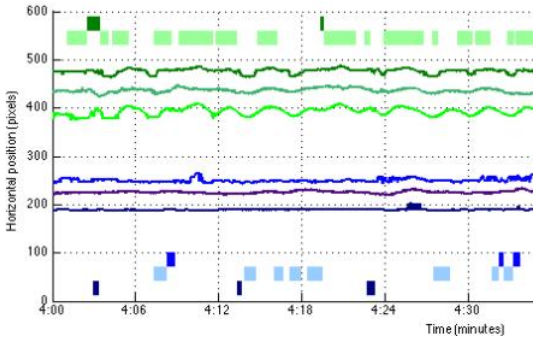


Figure 3 Speech and gesture activities for the left speaker (above) and right speaker (below).

The correlation between speech and gestures is strong, and can be seen in the correlation table, with about 62% of the participants' speech and gestures being in synchrony.

5.2 Movement patterns

Applying the video analysis method to the MINT dataset, we can also get interesting results related to various movement patterns and interaction synchrony among the conversation partners. As shown in Vels and Jokinen (2015b), a variety of gestures can be recognized by their combined movement curves, i.e. it is possible to recognize certain type of gesturing based on their characteristic bounding box trajectories. For instance, Figure 4 exemplifies beat gesturing, i.e. rhythmic hand gesturing during one's own speech, and clearly shows the variation in the front coordinates of the bounding box corresponding to hand movements. Figure 5 shows how a whole body moving forward provides a simultaneous set of back- and forward pikes in the curves related to upper body (hands) and lower body. Figure 6 shows how a large spike in the back coordinate of the body bounding box without movement in the head or the front coordinates of the body bounding box imply that there is gesturing behind the speaker. Finally, Figure 7 shows that if there are spikes both in the front and back coordinates of the body bounding box but the head coordinates is unchanged, the speaker waves her hands around.

6 Conclusion

We have discussed automatic recognition of human body movement and its use in communication studies. We used our previous algorithm for detecting body movement on video films and especially the version that can distinguish the three parts of the body: the head, the torso and the legs. We applied top-down-bottom-up approach and confirm earlier hypothesis of the gesturing and body movement as activities closely related to speaking.

The results show that the method can be applied with fairly good results, and combining the movement with speech occurrences we can visualise the interaction and especially the synchrony with speech and gestures. This is analogous to Campbell and Scherer (2010) who measured synchrony and alignment in spoken interactions, or Jokinen (2009) who applied the same method to measure conversational activity.

Future work concerns more detailed analysis of the MINT dataset and improving the hand and head movement detection algorithm. We will also use the same algorithm on other corpora and compare the gesturing in the context of intercultural communication. From the detected

body movements it is also interesting to try to extract gestures and their interpretation automatically. It is expected that the research presented in this paper can be used to integrate the user's body movement with the autonomous agent's gesture recognition capability, so as to produce natural interaction, and the models built using the help of bounded boxes and their visualisation as graphs will help to design the agent's own gesture model to produce appropriate gesturing and gesticulation in the course of the interaction.

Acknowledgement

The work has been done within the Estonian Science Foundation project MINT *Multimodal Interaction* (ETF 8958) and the project IUT 20-56 *Computational Models of Estonian*. I wish to thank Heiki-Jaan Kaalep for support, and Martin Vels for implementing the recognition algorithm.

References

Allwood, J. 2001. Dialog Coding—Function and Grammar. Gothenburg Papers. Theoretical Linguistics, 85. Department of Linguistics, Gothenburg University.

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta and P. Paggio. 2007. The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In Martin, J.C. et al. (eds.), Multimodal Corpora for Modelling Human Multimodal Behaviour. Special issue of the International Journal of Language.

Argyle, M. 1988. Bodily Communication. London: Methuen.

Argyle, M., Cook, M. 1976. Gaze and Mutual Gaze. Cambridge University Press.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford University Press.

Battersby, S. 2011. Moving Together: the organization of Non-verbal cues during multiparty conversation. PhD Thesis, Queen Mary, University of London.

Campbell, N., Scherer, S. 2010. Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity. *Proceedings of Interspeech*. Makuhari, Japan

Canny, J. 1986. A Computational Approach to Edge Detection, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6), pp. 679-698.

Endrass, B., Rehm M., Andre, E. 2009. Culture-specific communication management for virtual agents. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09). 281–288

Gonzales, R. C., Woods, R. E. 2010. *Digital Image Processing* (3rd edition). Pearson Education, pp. 652-661.

Jokinen, K. 2009. Gestures in Alignment and Conversation Activity. *Proceedings of the Conference of the Pacific Association for Computational Linguistics Conference (PACLING),* Sapporo, Japan, pp. 141-146.

Jokinen, K. 2009. *Constructive Dialogue Modelling:Rational Agents and Speech Interfaces*. Chichester: John Wiley.

Jokinen, K., Pelachaud, C. 2013. From Annotation to Multimodal Behavior. In Rojc, M. & Campbell, N. (Eds.) Co-verbal Synchrony in Human-Machine Interaction. CRC Press, Taylor & Francis Group, New York.

Jokinen, K., Tenjes, S. 2012. Investigating Engagement – Intercultural and Technological Aspects of the Collection, Analysis, and Use of Estonian Multiparty Conversational Video Data. *Proceedings of LREC'12*, pp. 2764 – 2769. Istanbul, Turkey: ELRA.

Kendon, A. 2004. *Gesture: Visible action as utterance.* New York: Cambridge University Press.

Kita, S. 2000. How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture*. pp. 162-185. Cambridge: Cambridge University Press. Open Access accepted version: http://wrap.warwick.ac.uk/66257/

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., Navarretta, C. 2010. The NOMCO Multimodal Nordic Resource – Goals and Characteristics. In *Proceedings of LREC'10,* Valetta, Malta: ELRA.

Schuller, B, Batliner, A. 2013. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley, 2013.

Vels, M., Jokinen, K. 2015a. Recognition of Human Body Movements for Studying Engagement in Conversational Video Files. In: Jokinen, K. & Vels, M. (eds) Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014. Tartu, Estonia. 110:014. Linkping: LiU Eletronic Press.

Vels, M., Jokinen, K. 2015b. Detecting Body, Head, and Speech in Engagement. *Proceedings of the IVA 2015 Workshop on Engagement in Social Intelligent Virtual Agents* (*ESIVA 2015).*

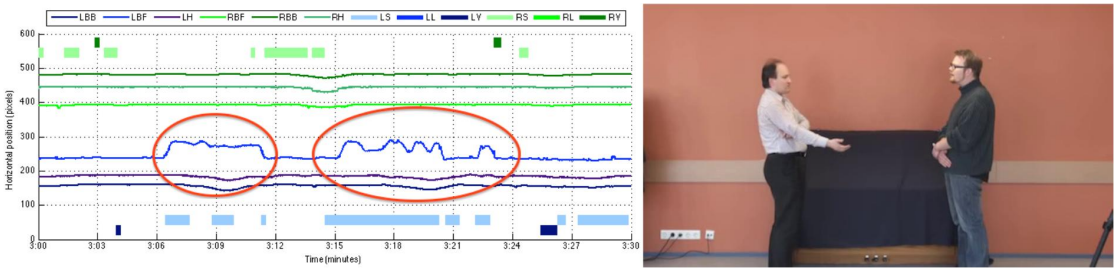


Figure 4 Beat gesturing simultaneous with speech (indicated by a light blue bar underneath the movement).

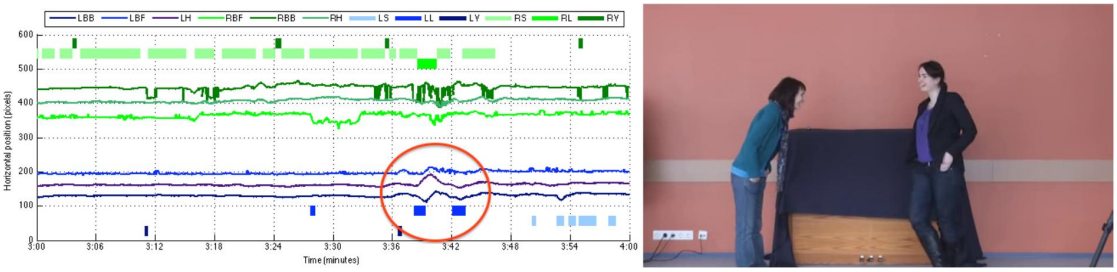


Figure 5 Leaning body movements in between speech and simultaneous with the partner's speech and gesturing.

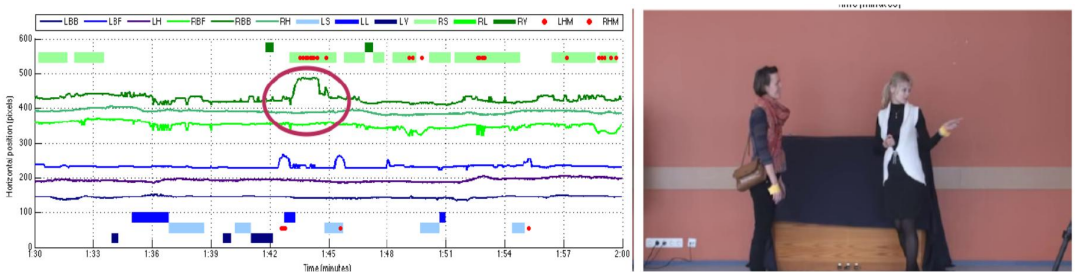


Figure 6 Large spike in the RBB coordinate without the RH or RBF => gesture somewhere behind the speaker.

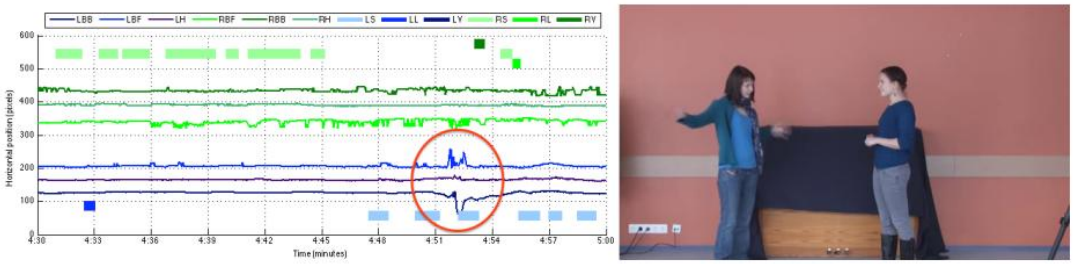


Figure 7 Spikes in both LBR and LBF coordinates with unchanged LH coordinate => the speaker waves her hands around

Multimodal perception in infants with and without risk for autism: A meta-analysis

Itziar Lozano [1], *Ruth Campos* [1], *Mercedes Belinchón* [1]

[1] Universidad Autónoma de Madrid, Department of Basic Psychology, (Spain)

itziar.lozano@uam.es, ruth.campos@uam.es, mercedes.belinchon@uam.es

Abstract

This manuscript shortly introduces a methodological proposal regarding how human beings process multimodal information at early ages in life. It specifically examines to what extent different developmental scenarios may lead to different trajectories of this capacity. Infants at genetic risk for Autism Spectrum Disorder (sibs-ASD, from now on) are at higher risk than the general population for either presenting ASD or showing subclinical traits. Due, firstly, to the genetic liability of the disorder and, secondly, to available evidence showing multimodal processing impairments in diagnosed individuals at several ages, sibs-ASD seem to represent an atypical situation of development particularly interesting to track the course of the ability. Although there is a lack of evidence explicitly exploring it in that sample as well as in ASD on the whole, many recently published works use multimodal stimuli to test how sibs-ASD process social information. Here we propose a meta-analysis that, although it is still unconcluded, aims to reorganize indirect evidence and shed light not only on the early development of multimodal processing in siblings-ASD, but also on methodological and theoretical issues related to the study of this crucial human ability.

Index Terms: multimodal processing, infants at risk for ASD, meta-analysis.

1. What is multimodal processing?

Multimodal information is highly present in daily life perceptual experiences. Generally, it refers to events containing inputs from different sensorial modalities. Socio-communicative situations have been considered of special interest for being highly redundant contexts (Pons and Lewkowicz, 2014; Bahrick, 2012). A typical early interaction scenario with parents simultaneously smiling, staring and touching their infants exemplifies a situation where auditory, touch and visual inputs co-occur. Despite infants are exposed to complex information concerning several sensory systems at the same time, they do not perceive it separately; instead, they perceive multimodal information as a coherent and meaningful unitary event (Bahrick, 2004).

Infants are exposed to a complex input from birth and their cognitive systems become progressively expert in processing multimodal input throughout development rather than efficiently processing it from early on. Accordingly, some authors have claimed that specialization mechanisms, where experience plays a key role, are likely to be involved in the development of the ability to process multimodal events (Lewkowicz and Ghazanfar, 2009; see Pons, Lewkowicz, Soto-Faraco and Sebastián-Gallés, 2009), similar to that in unimodal processing (see, for instance, Pons and Bosch, 2010 for an example on speech domain and Pascalis, de Haan and Nelson, 2002 on faces).

Becoming specialized on those socially relevant inputs allows for the development of crucial human functions later on—i.e., the use of symbols, language and mentalization, or object and event perception— (Lewkowicz, 2014; Bahrick, Lickliter and Flom, 2004). Thus, exploring how multisensory perceptual develops would allow to better understand perceptual, cognitive, and social development. Notice that, while those many sensorial modalities are involved in social interactions, we only focus on studying audiovisual redundancy due to it is highly present in communicative contexts. Consequently, from now on, we will be referring to the audiovisual sensory combination when using 'multimodal information processing' (MMP).

2. Multimodal processing and developmental trajectories

Some theoretical approaches (such as Neuroconstructivism) have highlighted developmental trajectories as a methodological alternative to explore typical and atypical development, mainly because they allow to track the course of human abilities over time (Thomas, Annaz, Ansari, Scerif, Jarrold and Karmiloff-Smith, 2009). Considering that MMP 1) starts evolving in the first days of life but changes are supposed to continue beyond adulthood and 2) is a dynamic phenomenon that depends both on experience and increasing abilities of infants, it seems that running this type of analysis could allow to trace changes over time in MMP performance. By testing infants at different early ages it is possible to draw how this ability differentially evolves leading to specialization courses in the ability that may vary among infants. The model assumes that there are as many different possible performances for the same ability as individuals. MMP exemplifies this variability, since some authors have claimed that there are some atypical neurodevelopmental scenarios where the ability does not evolve successfully. Exploring alternative trajectories may lead to know deeper about the underlying mechanisms involved in MMP development.

3. Why does it interest to explore multimodal processing in atypically developing scenarios?

Based on clinical and research evidences, we suspect that ASD (Autism Spectrum Disorder) may be one atypically developing scenario of special interest to study the early

development of MMP. ASD is a neurodevelopmental disorder of growing prevalence affecting near one percent of the population (CDC, 2007). From a clinical approach, it is defined by social communication impairments and restricted, repetitive patterns of behavior (American Psychiatric Association, 2013). However, affected individuals also show impairments in other abilities, such as Theory of Mind, executive function or central coherence, and unique processing styles which have been referred as the "cognitive phenotype of autism". It is now generally recognised that first-degree-relatives of affected individuals—mainly parents and siblings—are at increased risk for presenting sub-clinical forms of the clinical symptoms or the cognitive phenotype defining the condition, what is known as "broad autism phenotype" (Yirmiya and Ozonoff, 2007).

Particularly, infants who are siblings of older children with Autism Spectrum Disorder (siblings-ASD, from now onwards) present a heritable increased risk of developing the same disorder compared to the typical population, with a recurrence rate close to 19% (Ozonoff et al., 2011). Otherwise, they may show patterns of "broad autism phenotype" and milder impairments in abilities non-related with the core symptoms, which could lead them to differ from infants at low risk for autism in development (Yirmiya et al., 2006).

Some recent works suggest that MMP is impaired in individuals with autism at several ages—adults, adolescents and children— (Bebko, Weiss, Demark and Gomez, 2006; Massaro and Bosseler, 2003; Stevenson et al., 2014), although there is still no agreement on whether the lack of this ability could be considered as a characteristic of cognitive phenotype of ASD**.** Considering the genetic liability of the disorder along with these results showing that MMP is atypical in individuals with ASD at several ages, siblings-ASD may be an optimal sample to explore how this ability does develop in a non-normotypical scenario. Thus, when testing siblings-ASD in multimodal processing tasks it would be likely to expect that their developmental trajectories deviates from those followed by typically developing infants.

4. Evidence on multimodal processing in siblings-ASD

Some recent theoretical works have highlighted the interest of studying the development of MMP in atypical scenarios—and, especially ASD— (Bahrick, 2010; Bahrick and Todd, 2012; Hill, Crane and Bremner, 2012). Given the agreement on the need for exploring this phenomenon, we aimed to exhaustively search for evidence on the early development of the MMP in siblings-ASD. When doing a systematic search, a paradoxical result arose: although many studies explore social competence in siblings-ASD by using multimodal stimuli in tasks —probably due to that multimodal information is more salient in social interaction contexts—, there is still comparatively little published research on siblings-ASD, or even on ASD on the whole about this specific topic. In fact, we found a single work explicitly focused on the study of the ability itself in these groups (see Guiraud et al., 2012). That finding seems to point out that to date MMP has been only indirectly explored, which could partially explain the methodological heterogeneity found in aims, paradigms, methods and many other methodological issues, as well as the lack of direct unimodal and multimodal comparisons within a single experiment design. Making an effort to organize such heterogeneity may clarify firstly what variables could explain the results of previous (indirect) research on how siblings-ASD process multimodal information and, secondly, what variables (regarding participants, tasks, materials and other methodological aspects) could be more relevant to theoretically explain the process itself as well as the expected trajectories differences. But could these data be systemized?

5. Meta-analysis as an alternative to reorganise the study of multimodal processing in siblings-ASD

One tool seems particular suitable for that aim. Meta-analysis is a quantitative procedure that arose as an alternative of narrative and systematic reviews. It came out as a step beyond those two owing to it allows describing, integrating and analysing empirical data of primary studies regarding a specific research topic.

Meta-analysis is commonly defined by (Botella and Gambara, 2002):

1) Being precise, since it requests information about specific questions.

2) Being able to measure numerically to what extent data support these questions.

3) Replicating, as any other researcher would repeat by following the same steps and finally obtain similar results.

Among the advantages that it offers, its main potential contribution to our aim is probably that it allows to define organisers not included in the data of the primary studies selected. In other words, despite the lack of results explicitly exploring the ability, measures on MMP from studies exploring social performing in siblings-ASD may be reorganized under a new theoretical analysis different from those supported by primary studies.

According to Botella and Gambara (2006), although meta-analysis do not necessarily follow a linear sequence of stages, the tasks involved in the procedure would follow a logical order starting from 1) defining the problem through operations and hypothesis; 2) doing the search; 3) categorising the studies; 4) transforming data to a common metric; and, finally, 5) analysing and 6) discussing them. The following sections describe steps that have been completed by the time we are writing this manuscript (that is, from 1 to 3).

5.1. Defining questions and variables

One of the first steps —and, possibly, one of the hardest when running a meta-analysis— is to outline relevant questions and the associated variables that may allow to explore them. Considering our aims, we drew the key questions shown in Table 1.

Table 1. *Questions and variables*

QUESTIONS	VARIABLES
1. *Do infants at risk for ASD show difficulties when processing social information?*	**V1:** Multimodal Processing Performance
2. *Are these impairments modulated by sensory modality?*	**V2:** Sensory modality
3. *What is the developmental pattern*	**V3:** Age

of this ability?	
***4.** Do those impairments change with age depending on sensory modality?*	**V4:** Age x sensory modality

To answer the first question, it is necessary to extract from the studies 'the mean performance in tasks using multimodal stimuli' of infants at high and low risk for ASD, in order to empirically compare them as groups.

As it has been pointed out, there are no previous studies explicitly comparing the influence of sensory modality (unimodal *vs*. multimodal) in a complete experimental design. However, studies exploring social development in siblings-ASD use either unimodal or multimodal tasks. For that reason, we also decided to codify the modality of the stimuli used in the task of studies exploring social processing in order to indirectly test its possible influence in group differences (as question 2 indicates).

Questions 3 and 4 refer to whether impairments in social processing would change throughout development and, if they do, whether it varies depending on the uni or multimodal nature of stimuli.

5.2. Location of the studies

Meta-analysis requires a search strategy that needs to be developed with care. Once key questions and variables were defined, it was time to decide what keywords would be included. Doing the search mainly involves to enclose the main representative results arising from the questions and variables defined previously and, at the same time, to exclude papers that do not meet certain criteria. Regarding keywords, 'at risk' and 'siblings' as well as 'ASD' and 'autism' were finally selected (both as synonymous pairs) for representing the sample. A full search on PsycINFO and PsycARTICLES—frequently used databases in the concerned topic—was conducted in August 2015. We aimed to run a search as inclusive as possible by combining the following keywords and *Boolean* operators: ('high risk' OR 'siblings') AND ('autism' OR 'ASD'). In addition, results were restricted by age by entering only works that assessed infants, toddlers and children. Despite our search focused on measures on MMP from studies exploring social performing in siblings-ASD, we decided not restricting the search to studies exploring social performance because many works did not explicitly state that term as a keyword or in the abstract although they explored it (for instance, studies exploring processing of faces or speech). Finally, we indicated that duplicates were excluded from the final sample.

5.3. Inclusion criteria

By using all the restrictors mentioned we obtained a search containing 1199 studies that was narrowed down by the following inclusion criteria (summarized in Table 2) that filtered studies not fitting with our key questions: 1) The studies mainly focused on exploring social processing; 2) at least two groups of infants were tested (high and low genetic risk for ASD) ; 3) participants aged from 0 to 36 months old (that is, from birth to age of diagnosis); 4) the studies run experimental designs where conditions are carefully manipulated (remaining excluded, for instance, theoretical reviews or observational studies); 5) quantitative continuous measures were registered (namely, reaction or fixation times and latencies). Furthermore, only studies showing all the information needed for future calculation were incorporated (either descriptive or contrast statistics). Any study not meeting these criteria were excluded. The final selection included 47 academic journal articles as well as dissertations and posters. It is also worth noting that, among that sample, 18 were collected from secondary informal sources such as the references cited in the papers primarily found as well from the main research groups' websites (namely, BASIS Team, Autism Speaks and BSRC) as well as those of its members. In both cases, we aimed to avoid selecting biased results, a trend that is usually referred as 'publication bias'.

Table 2. *Exclusion criteria*

Domain	Social
Comparison groups	Risk for ASD (high and low)
Age	0-36 months
Design type	Experimental
Measures	Quantitative continuous

5.4. Coding of characteristics

After doing the search we defined moderator variables, that is, those considered more likely to play a role in the effect sizes arisen from the expected high-risk and low-risk between-groups' differences. Thus, our selection was newly restricted by questions and variables shown in Table 1. We report those considered relevant to our aims in Table 3.

The only relevant participant characteristic for our analysis was infants' 'Age'. We entered mean age in days for each group of participants, which may contribute to answer Questions 3 and 4 shown in Table 1 (*What is the developmental pattern of this ability?* and *Does those impairments change with age depending on sensory modality?)*.

The remaining coded variables mostly involved features regarding stimulus, namely: 'type of modality', defined as a dichotomy between unimodal or multimodal depending on whether information shown in the task belong to one or more sensory modalities; 'sensory modality', that is, whether the stimuli used in the tasks contained neither auditory, visual nor audiovisual input); 'sensory dominance', which only refers to those tasks that include audiovisual multimodal with a predominance of either visual or auditory information; 'stimuli content', that refers to the nature of the information contained in the task, which goes from faces or objects to speech or non-speech sounds; and, finally, 'other stimuli features' (related to whether they are static or dynamic, simple or complex, etc.). We decided to codify those variables for being at the core of our main hypothesis claiming that sensory nature of the stimuli may influence on how siblings-ASD process social information (corresponding to Questions 1 and 2 in Table 1), which probably differs throughout development (as Question 3 points out), meaning that both type of variables (regarding the nature of the stimuli and age) presumably mutually interact (see Question 4).

We also coded detailed information regarding methodological issues, such as the dependent variables measured (for instance, fixation times, reaction times or

latencies) as well as the instruments (whereas some studies use *eye-tracker* others record EEG) and methods used (such as habituation or head preference), since we hypothesize that they may play a role in the direction of between-groups differences. Finally, we decided to codify variables—such as year and author— that sometimes show unexpected results.

Table 3. *Variables codified (with some examples in brackets)*

Age (Mean group age, in days)
Type of modality (Unimodal/Multimodal)
Sensory dominance (Audio-visual/Visual-auditory)
Sensory modality (Auditory/Visual/Audiovisual)
Stimuli Content (Faces, sounds, speech, objects, etc.)
Other Stimuli Features (Static/Dynamic, Simple/Complex)
Instrument (*Eye-tracker*, EEG, HHP, NIRS, etc.)
Dependent Variable (Fixation time and %, amplitude and latencies, visual preference, reaction time, latencies etc.)
Paradigm (Habituation, head preference, EEG, NIRS. etc.)
Year
Author

6. Conclusions

Meta-analysis is a methodological tool that may help to reinterpret data more systematically by detecting the relationship between mediating variables, such as the possible role of sensory modality of stimuli and age, in how infants at higher risk for ASD process socially relevant information. Based on the progress made so far in our research, we believe that both the results and the process itself of this meta-analysis will allow us to identify cues that could be relevant not only on the early development of MMP in siblings-ASD, but also on methodological and theoretical issues related to the study of this crucial human ability.

7. Acknowledgements

This study has been supported by the Spanish Ministry of Education under the FPU Predoctoral Grant program (Reference: FPU13/03508).

8. References

[1] F. Pons and D.J. Lewkowicz, "Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech", *Acta Psychologica*, vol. 149, pp. 142–147, 2014.

[2] L. E. Bahrick and J. T. Todd, "Multisensory processing in autism spectrum disorders: Intersensory processing disturbance as a basis for atypical development", in *The New Handbook of Multisensory Processes*, B.E., Stein, Cambridge, MA: MIT Press, 2012, ch.40, pp. 657-674.

[3] L.E. Bahrick. "The development of perception in a multimodal environment", in *Theories of Infant Development*, G. Bremner, and A. Slater, Malden, MA: Blackwell Publishing, 2004, ch. 4, pp. 90-120.

[4] L.E. Bahrick, R. Lickliter and R. Flom. "Intersensory redundancy guides infants' selective attention, perceptual and cognitive development". *Current Directions in Psychological Science*, vol. 13, pp. 99-102, 2004.

[5] D.J. Lewkowicz and A.A. Ghazanfar. "The emergence of multisensory systems through perceptual narrowing". *Trends in Cognitive Sciences*, vol. 13, no. 11, pp. 470–8, 2009.

[6] F. Pons, D.J. Lewkowicz, S. Soto-Faraco and N. Sebastián-Gallés. "Narrowing of intersensory speech perception in infancy". *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10598-10602, 2009.

[7] F. Pons and L. Bosch. "Stress pattern preference in Spanish-learning infants: The role of syllable weight". *Infancy*, vol. 15, no. 3, pp. 223-245, 2010.

[8] O. Pascalis, M. de Haan and C.A. Nelson. "Is face processing species-specific during the first year of life?" *Science*, vol. 296, no. 5571, pp. 1321–1323, 2002.

[9] D.J. Lewkowicz. "Early Experience and Multisensory Perceptual Narrowing". *Developmental Psychobiology*, vol. 56, no. 2, pp. 292–315, 2014.

[10] L.E. Bahrick, R. Lickliter and R. Flom. "Intersensory Redundancy Guides the Development of Selective Attention, Perception, and Cognition in Infancy". *Current Directions in Psychological Science*, vol. 13, pp. 99–102, 2004.

[11] M.S. Thomas, D. Annaz, D. Ansari, G. Scerif, C. Jarrold and A. Karmiloff-Smith. "Using developmental trajectories to understand developmental disorders". *Journal of speech, language, and hearing research*, vol. 52, no. 2, pp. 336-358, 2009.

[12] CDC (Center for Disease Control and Prevention-Atlanta). *Prevalence of Autism Spectrum Disorders*. *Autism and Developmental Disabilities Monitoring Network*. Morbidity and Mortality Weekly Report (MMWR) 56, SS-1, 2007.

[13] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, DC: American Psychiatric Association, 2013.

[14] N. Yirmiya and S. Ozonoff. "The Very Early Autism Phenotype". *Journal of Autism and Developmental Disorders*, vol. 37, no. 1, pp. 1–11, 2007.

[15] S. Ozonoff, G.S. Young, A. Carter, D. Messinger, N. Yirmiya L.Zwaigenbaum, … and W.L. Stone. "Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study". *Pediatrics*, vol. 128, no. 3, pp. e488-e495, 2011.

[16] N. Yirmiya, I. Gamliel, T. Pilowsky, R. Feldman, S. Baron-Cohen and M. Sigman. "The development of siblings of children with autism at 4 and 14 months: Social engagement, communication, and cognition". *Journal of Child Psychology and Psychiatry*, vol. 47, no. 5, pp. 511-523, 2006.

[17] J.M. Bebko, J.A. Weiss, J.L. Demark and P. Gomez. "Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism". *Journal of Child Psychology and Psychiatry*, vol. 47, no. 1, pp. 88-98, 2006.

[18] D.W. Massaro and A. Bosseler. "Perceiving speech by ear and eye: Multimodal integration by children with autism". *Journal of Developmental and Learning Disorders*, vol. 7, pp. 111-144, 2003.

[19] R.A. Stevenson, J.K. Siemann, B.C. Schneider, H. E. Eberly, , T.G. Woynaroski, S.M, Camarata and M.T. Wallace. "Multisensory temporal integration in autism spectrum disorders". *The Journal of Neuroscience*, vol. 34, no. 3, pp. 691-697, 2014.

[20] L.E. Bahrick. "Intermodal perception and selective attention to intersensory redundancy: Implications for typical social development and autism", in *Wiley-Blackwell Handbook of Infant Development*, J.G. Bremner, T.D. Wachs, vol. 2, Oxford, UK: Wiley-Blackwell, 2010, ch. 4, pp. 120-165.

[21] L.E. Bahrick and J.T. Todd. *Multisensory processing in autism spectrum disorders: Intersensory processing disturbance as a basis for atypical development*, in *The New Handbook of Multisensory Processes*, G.A. Calvert, Ch. Spence and B.E. Stein, Cambridge, MA: MIT Press, 2012, ch. 40, pp. 657-674.

[22] E.L. Hill, L. Crane and A.J. Bremner. *Developmental disorders and multisensory perception*, in *Multisensory Development*, A.J. Bremner, D.J. Lewkowicz, Ch. Spence, Oxford: Oxford University Press, 2012, ch. 12, pp. 273-300.

[23] J. Guiraud, P. Tomalski, E. Kushnerenko, H. Ribeiro, K. Davies, T. Charman, … M. H. Johnson, "Atypical audiovisual speech integration in infants at risk for autism". *PloS One*, vol. 7, no. 5, pp. e36428, 2012.

[24] J. Botella and H. Gambara. *¿Qué es el meta-análisis?* Madrid: Biblioteca Nueva, 2002, pp. 28-29.

[25] J. Botella and H. Gambara. "Doing and reporting a meta-analysis". *International Journal of Clinical and Health Psychology*, vol. 6, no. 2, pp. 425-440, 2006.

Annotation and Multimodal Perception of Attitudes: A Study on Video Blogs

Noor Alhusna Madzlan[1 3], *Justine Reverdy*[2], *Francesca Bonin* [2],
Loredana Cerrato[2], *Nick Campbell*[2]

[1] CLCS, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin
[2] SCSS, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] ELLD, Faculty of Languages and Communication, UPSI, Malaysia

madzlann@tcd.ie, reverdyj@tcd.ie, boninf@tcd.ie, cerratol@tcd.ie, nick@tcd.ie

Abstract

We report the set-up and results of an experiment designed to verify to what extent attitudes can be identified and labelled by using an *ad hoc* annotation scheme. Respondents were asked to label the multimodal expressions of attitudes of a number of video bloggers selected from a vlog corpus. This study aims at measuring respondents' attitude choice as well as the difference in their attitude judgments. We investigate the contribution of different modalities to the process of attitude choice (audio, video, all). The results are analysed from three perspectives: inter-annotator agreement, contribution level for each modality and certainty level of attitude choice. Participants showed to perform better in perceiving attitudes when they were presented with the combined audio-visual stimuli in comparison to the audio only and video only stimuli. Participants showed to be more certain in selecting "Friendliness" than the other attitudes.
Index Terms: multimodal perception, video blogs, annotation,affective states.

1. Introduction

Communicative content in human communication involves the expression of social attitudes, defined as social affective states that the speakers intend to transmit to the audience as defined in [1]. Differently from emotions, attitudes might not correspond to the truth inner psychological state of the speaker, but represent what the speaker intentionally wants to show to the outside. Understanding how speakers express their social attitudes is a fundamental step in the process of successful communication both in human-human and human-machine interactions. While many researchers focus on detection of emotion in human-human conversations [2][3][4][5][6], less attention has been given to the analysis of social attitudes.

Nevertheless, understanding how speakers express their attitude by means of different verbal and visual feature is essential to establishing a successful communication and it is particularly useful when it comes to implementing better systems for Human Machine Interactions and Human Robot Interactions, because it can provide the machine with knowledge related to the socio-affective states of the participant.

Understanding of the rich communication content in terms of social signals provides invaluable skills in technologies such as companion systems, socially aware interaction systems, conversational agents. [7].

Previous studies in the field of Linguistics, Social Signal Processing and Affective Computing have highlighted the importance of integrating the information carried out by social signals, in particular emotions, affective states and attitudes in the process of analysing and interpreting the communicative content of interactions [8][9][10]. In this study we consider a specific communicative situation: video blogs (VLOGS) where speakers tend to have a dynamic representation of attitude expression in a specific scenario of social interaction. We focus our attention on how to define and label attitude expressions in a corpus of video blogs selected from Youtube. In order to label attitudes we defined an annotation scheme to annotate the vlog corpus. Our annotation scheme, named N5, is a derivation of the standard A10 attitude annotation proposed by Henrichsen and Allwood [11].

In this paper we present the results of an experiment in which we asked respondents to label multimodal expressions attitude of video bloggers. The aim of this study is to see to what extent attitudes can be identified and labelled by using our *ad hoc* annotation scheme.

2. Related Work

Recent studies explore communicative content, which includes affect and attitudes with its relation to their perceptual meanings [12][13] [14] [15].

Morlec et al. [12] suggest that attitudes strongly reflect in the prosody of the speaker. Their study introduced six attitudes expressed in French from the inter-perceptual-center group (IPCG) melodic curve corpus, which consist of 322 utterances for each of the six attitudes, which are Assertion, Question, Exclamation, Incredulous Question, Suspicious Irony and Evidence. They conducted a perception study among 20 participants to validate the six attitudes using training and testing sentence modules. Results suggest that there exist confusions between Incredulous, Question and Suspicious Irony despite clear prosodic distinctions.

Rilliard et al. [14] conduct a perceptual study of the prosodic characteristics of attitudes (defined as prosodic attitudes) through audio-visual modalities. Extending work on six prosodic attitudes developed by Morlec et al. [12], they included audio-visual recording of the six attitudes from two French speakers and developed a perception test to present different modalities to 32 French listeners. Results show that the Audio-Visual modality prove most helpful for listeners to identify these prosodic attitudes, particularly Obviousness and Suspicious Irony. Despite attaining good recognition rates for each of the 6 attitudes, an interesting approach of analysis is the application of a cluster analysis to understand confusions between these attitudes. Analysis found that Doubt-Incredulity and Surprise-Exclamation are confused in the audio modality, while Question and Doubt-Incredulity are confused when pre-

sented in video stimuli. For the audio-video stimuli, video helps in distinguishing Exclamation from Doubt-Incredulity. This work is helpful in providing clear distinctions between the six prosodic attitudes by conducting a perception study and cluster analysis through different modes of stimuli.

Similar to [14], Allwood et al. [16] conducted a perception study on attitude (defined here as Affective-Epistemic States (AES)) using multimodal stimuli. The study involves 12 Swedish participants presented with recordings from the NOMCO First Encounter. Gestures are annotated based on the MUMIN annotation scheme [17]. Participants are shown a two-minute long clip of the corpus and are required to choose any words that describe both affective-epistemic and behavioural states. Results from semantic analysis lead to seven types of AES: happiness, interest, nervousness, confidence, disinterest, thoughtfulness and understanding. Audio-visual modality shows most attributions for nervousness, interest and thoughtfulness. Further analysis suggests that AES expression may be conflicting or complementing according to different modalities. Happiness, for instance, is expressed best through the audio modality but not vividly shown in video modality. This finding claims that multimodal expressions of AES are more complex to perceive.

Findings from these research works suggest that perception studies, through several methods, are typically used to validate the choices of attitudes. With reference to [16], our work elaborates on a similar method of validating our attitude choices through an online perception study.

3. N5 Attitude Categories

Our past work on developing an attitude recognition system [18][19] conducts data annotation using an adaptation of an attitude annotation scheme derived by Henrichsen and Allwood [11]. This annotation scheme consists of ten attitudes named A10, as listed in Table 1.

A10	
Amused	Bored
Casual	Confident
Enthusiastic	Friendly
Impatient	Interested
Thoughtful	Uninterested

Table 1: *Standard A10-based Annotation Scheme*

On the basis of A10, we developed a new annotation scheme, hereafter, N5, constituted by 5 categories, presented in Figure 1. Our hypothesis is that those categories are more representative of the attitudes present in our corpus of video blog. Four of the categories in our N5 annotation scheme are taken from the A10 annotation scheme and the category "Frustration" is added because it was considered to be appropriate for our vlog corpus.

In order to validate our hypothesis, we asked two Linguist experts to annotate a total of 250 vlogs [19] using the N5 scheme. We then calculated their inter-rater agreement, which resulted to 0.75 Cohen's Kappa. The reasonably high Kappa shows how the 5 categories are a good representation of the attitude in the corpus. However, in order to have a further validation, we also run a perceptual test involving a group of anonymous non-expert public participants.

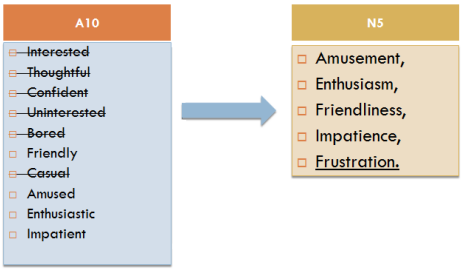


Figure 1: *The N5 attitude categories*

4. Perceptual Test Setup

We designed and run a perceptual test with three different aims:

i.) Validate the choice of the five attitude categories in the N5 annotation scheme
ii.) Investigate which of the modality (audio,video,combined) mostly contributes to the attitude selection task
iii.) Investigate the certainty level of the participants

Twenty participants, recruited among Trinity College Dublin (TCD) staff and students, took part anonymously in the experiment on a voluntary basis. They were requested to provide age and gender information and to read participation information before starting the test. A clearance from the SCSS Research Ethics Committee was obtained previous to the study.

Participants were provided with a link to an online survey and were given 20 minutes to answer all questions. The online survey was developed in-house using PHP5 with an MVC architecture associated with a MySQL database.

The test consisted of three phases. In order to validate the N5 scheme, *i)*, participants were provided with N5 categories and had an additional choice showing the remaining categories from the A10 scheme listed under a drop-down menu with the headings "Others". The participants presented with the stimuli had to select one of the categories to describe the affective state.

In order to investigate which of the modality (audio,video, combined) mostly contributes in the attitude recognition task, *ii)*, participants had to label a total of 58 stimuli presented in three sections of 18 questions each. Section A consists of the audio only stimuli, Section B comprises video only stimuli (audio muted) and Section C presents both audio-video stimuli.

Finally, *iii)* to investigate the certainty level, after selecting an attitude, participants were asked to decide, on a scale ranging from 1 to 7 (going from Unsure to Very Certain), how certain they were about their judgments on their attitude selection. An example of this certainty scale is pictured in Figure 4.

5. Results

We analysed the results from three perspectives: inter-annotator agreement, contribution level for each modality and certainty level of attitude choice.

5.1. Inter-annotator agreement

Results achieved 100% agreement among all the participants for 37% of the stimuli. We further conducted inter-annotator agreement, and found a "fair agreement" between all 20 raters with a k-value of 0.27 using weighted Fleiss Kappa [20]. The

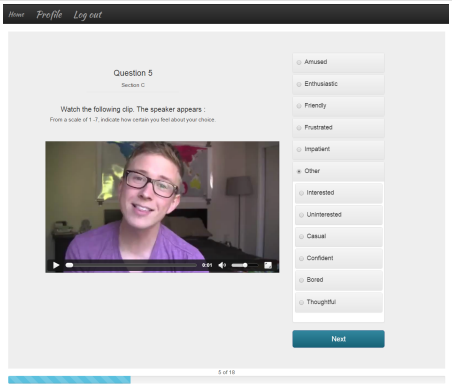
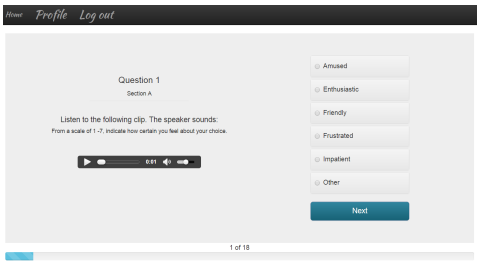
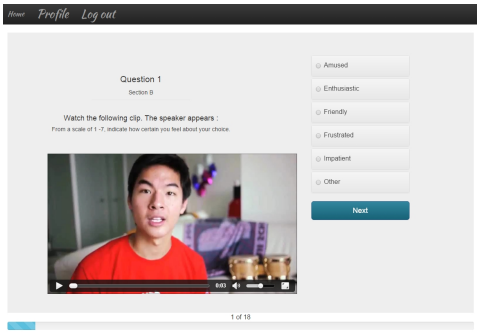


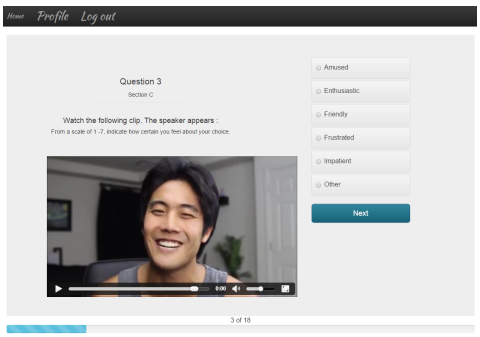
Figure 2: *N5 and "Other" attitude choices*



(a) *Section A*



(b) *Section B*



(c) *Section C*

Figure 3: *Examples of all sections*

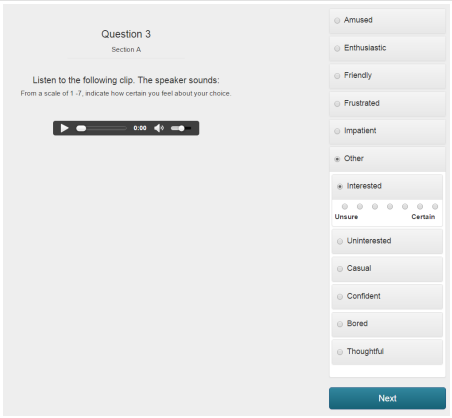


Figure 4: *Example of certainty scale*

low value for agreement is not surprising considering the large number of raters (20 raters) involved in the test. In general it is possible to observe that Frustration is often preferred over other attitudes (see Figure 5).

This observation is in agreement with our justification for the inclusion of the "Frustration" state as an attitude class that is salient in the vlog dataset. Figure 5 shows also that category "Other" did not get enough choices to justify inclusion in our N5.

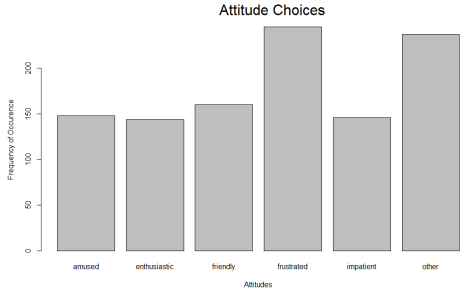


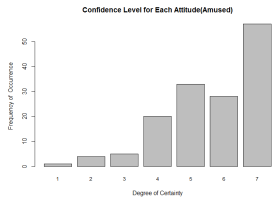
Figure 5: *Frequency of Occurrence of selected Attitudes*

5.2. Contribution of modality in the attitude selection task

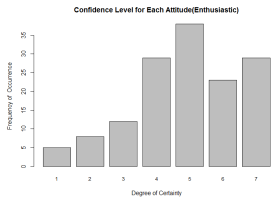
We further analysed the relevance of multimodalities for attitude perception and observed that a fusion of audio and visual information is most helpful for participants to perceive attitude expressions of vlog speakers. Specifically, annotators reached a precision of 35.5% when exposed to the Audio+Video stimulus, of the 33.1% while exposed to Audio only and of 31.6% while exposed to video only.

5.3. Certainty level for attitude choice

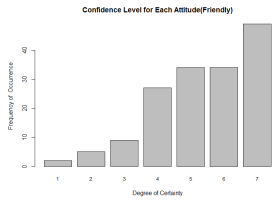
Following that, we conducted analysis on the certainty level of participants with their attitude choice. Figure 6 shows levels of certainty per attitude.
Participants showed to be most certain when selecting "Impatience" and "Friendliness", while they showed less certainty when selecting the categories listed under "Other".



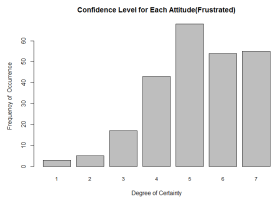
(a) Certainty Level for Amusement



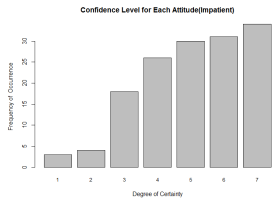
(b) Certainty Level for Enthusiasm



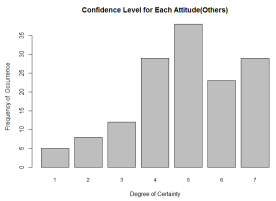
(c) Certainty Level for Friendliness



(d) Certainty Level for Frustration



(e) Certainty Level for Impatience



(f) Certainty Level for Other

Figure 6: Certainty levels for each attitude

This suggests that participants were not certain and most confused about their choice of the "Other" category.

6. Discussion and Conclusions

The main aim of this study was the validation of our novel attitude annotation scheme N5. To achieve this aim we performed a perception test with 20 participants who were asked to annotate a subset of our vlog corpus.

The low inter-annotator agreement is expected and in-line with Schuller's [21] statement on the difficulty in obtaining reliability in the annotation of affective states due to the equivocal nature of affect data. Factors like age, gender and cultural backgrounds of participants may contribute to this variation. This finding is not unexpected as it is challenging to assign labels to these kind of phenomena since attitude perception is subjective.

"Frustration" was chosen most of all 11 attitudes, making this a relevant label to annotate attitude in the vlog data. On the other hand, there was not sufficient consistency in the "Other" category to justify inclusion of an extra attitude. This findings suggest that our N5 attitude categories seems to be a sufficient scheme to annotate attitudes in our vlog corpus.

Participants showed to perform better in perceiving attitudes when they were presented with the audio-visual signals in comparison to the audio only and video only stimuli. We found that the fusion of multimodalities from the vlog data is in agreement with Shochi et al. [22], who also report that Audio-Visual modalities have stronger influence in attitude perception.

To further understand which attitude categories are clearly detected and which of the attitudes participants have reservations about, we conducted a certainty test. We notice that participants were more certain in selecting "Friendliness" to the other attitudes. Another observation from this measure of certainty is that the participants showed uncertainty when selecting the attitudes from the drop down menu Others. This is interesting for us as the attitudes included in the "Others are those from the A10 Attitude annotation scheme which we decided not include in the N5 scheme, as we assumed they were not represented in our vlog corpus. This level of uncertainty among participants may be an indication that the attitudes from the"Others list are indeed not so representative of our vlog corpus.

7. Future work

Our work presents a perception study to validate the choice of attitude categories in our vlog dataset. As an extension of this work, the application of this results will be implemented in a predictive classifier in developing a computational framework for automatic attitude recognition. To further improve the current findings, we suggest plausible methods for measuring attitude perception. Due to varying results from multi-rater agreement test, we plan to analyse confusion matrix and/or perform cluster analysis to explain these discrepancies. Future work is also planned for an in depth analysis of gender and age effects to better understand factors that can contribute to attitude perception.

8. Acknowledgements

This work is supported by the English Language and Literature Department, UPSI, Ministry of Education Malaysia, Center for Excellence for Digital Content Technology (ADAPT) at TCD and the Speech Communication Laboratory at TCD.

9. References

[1] Y. Lu, V. Aubergé, A. Rilliard *et al.*, "Do you hear my attitude? prosodic perception of social affects in mandarin," *Proceedings of Speech Prosody 2012*, pp. 685–688, 2012.

[2] P. Ekman, "Are there basic emotions?" 1992.

[3] A. Kappas and N. Krämer, *Face-to-Face Communication over the Internet: Emotions in a Web of Culture, Language, and Technology*, ser. Studies in Emotion and Social Interaction. Cambridge University Press, 2011. [Online]. Available: http://books.google.ie/books?id=ofM_AHampHsC

[4] J. A. Hall and D. Matsumoto, "Gender differences in judgments of multiple emotions from facial expressions." *Emotion*, vol. 4, no. 2, p. 201, 2004.

[5] C. Yoo, J. Park, and D. J. MacInnis, "Effects of store characteristics and in-store emotional experiences on store attitude," *Journal of Business Research*, vol. 42, no. 3, pp. 253–263, 1998.

[6] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[7] P. Persson, J. Laaksolahti, and P. Lönnqvist, "Understanding socially intelligent agents-a multilayered phenomenon," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 31, no. 5, pp. 349–360, 2001.

[8] A. Vinciarelli and G. Mohammadi, "Towards a technology of nonverbal communication: Vocal behavior in social and affective phenomena," igi-global, Tech. Rep., 2010.

[9] M. D. Pell, "Judging emotion and attitudes from prosody following brain damage," *Progress in brain research*, vol. 156, pp. 303–317, 2006.

[10] D. Ballin, M. Gillies, and I. Crabtree, "A framework for interpersonal attitude and non-verbal communication in improvisational visual media production," 2004.

[11] P. J. Henrichsen and J. Allwood, "Predicting the attitude flow in dialogue based on multi-modal speech cues," *NEALT PROCEEDINGS SERIES*, 2012.

[12] Y. Morlec, G. Bailly, and V. Aubergé, "Generating the prosody of attitudes," in *Intonation: Theory, Models and Applications*, 1997.

[13] J.-M. Blanc and P. F. Dominey, "Identification of prosodic attitudes by a temporal recurrent network," *Cognitive Brain Research*, vol. 17, no. 3, pp. 693–699, 2003.

[14] A. Rilliard, J.-C. Martin, V. Aubergé, T. Shochi *et al.*, "Perception of french audio-visual prosodic attitudes," *Speech Prosody, Campinas, Brasil*, 2008.

[15] D.-K. Mac, V. Aubergé, A. Rilliard, and E. Castelli, "Cross-cultural perception of vietnamese audio-visual prosodic attitudes," in *Speech Prosody*, 2010.

[16] J. Allwood, S. Lanzini, and E. Ahlsén, "Contributions of different modalities to the attribution of affective-epistemic states," in *Proceedings from the 1st European Symposium on Multimodal Communication University of Malta*, pp. 1–6.

[17] J. Allwood, L. Cerrato, L. Dybkjaer, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin multimodal coding scheme," in *Proc. Workshop on Multimodal Corpora and Annotation*, 2005.

[18] N. A. Madzlan, J. Han, F. Bonin, and N. Campbell, "Towards automatic recognition of attitudes: Prosodic analysis of video blogs," *Speech Prosody, Dublin, Ireland*, pp. 91–94, 2014.

[19] N. Madzlan, J. Han, F. Bonin, and N. Campbell, "Automatic recognition of attitudes in video blogs - prosodic and visual feature analysis," in *INTERSPEECH*, 2014.

[20] J. L. Fleiss, J. Cohen, and B. Everitt, "Large sample standard errors of kappa and weighted kappa." *Psychological Bulletin*, vol. 72, no. 5, p. 323, 1969.

[21] B. Schuller, "Multimodal affect databases: Collection, challenges, and chances," *The Oxford Handbook of Affective Computing*, pp. 323–333, 2014.

[22] T. Shochi, D. Erickson, A. Rilliard, V. Aubergé, J.-C. Martin *et al.*, "Recognition of japanese attitudes in audio-visual speech," in *Speech prosody*, vol. 2008, 2008, pp. 689–692.

The functions of fillers, filled pauses and co-occurring gestures in Danish dyadic conversations

Costanza Navarretta

University of Copenhagen
costanza@hum.ku.dk

Abstract

Fillers, alone or accompanied by pauses and/or gestures, are quite frequent in all types of spoken communication. They have numerous and non-exclusive functions which are related to interaction management (feedback and turn management) or discourse planning. Fillers are part of the language and thus, to some extent, language dependent. This article presents an analysis of fillers, filled pauses and co-occurring gestures in a Danish multimodal corpus of first encounters. The aims of the study are to determine the most common fillers in the corpus, the gestures co-occurring with them, their functions, and possibly their most prototypical uses. The results of our study indicate that the most common fillers in the data are *øh*, *mm*, *øhm* which all are accompanied by one or more gestures in most of their occurrences. We also found that each filler type has a predominant or prototypical use. *Mm* often occurs alone as feedback marker and is accompanied by feedback gestures. *Øhm* has the longest duration and often precedes an utterance or a clausal phrase signaling discourse planning. Its co-speech gestures have also interaction management functions. Finally, *øh* often precedes a content word, has a shorter duration than *øhm* and signals lexical retrieval. Interestingly the prototypical uses of the vocal *øh* and the vocal-nasal *øhm* are the same as those of the English vocal *uh* and vocal-nasal *um*, respectively.

Index Terms: multimodal communication, gestures, filled pause

1. Introduction

Face-to-face communication is multimodal including at least auditory (speech) and visual (gestures) modalities. The modalities are not only temporally, but also semantically related at many levels. This papers is about a particular phenomenon of face-to-face communication, the so-called fillers, such as the English *uh* and *um*. Fillers are very frequent in spoken language and can occur alone or in conjunction with a speech pause (filled pauses). The language accounted for is Danish. The paper also addresses the gestures which co-occur with the Danish fillers and their functions. The gestures included in this study are head movements, facial expressions, body postures and hand movements.

Fillers have multiple, non-exclusive functions which are related to interaction management [1, 2, 3] and cognitive processes of discourse planning and word retrieval [4, 5]. Researchers have noticed that there is an inverse frequency relation between hand gestures and filled pauses [6, 7] and that many hold gestures co-occur with filled pauses [8, 9].

Fillers are an integral part of the language and have therefore language specific characteristics [10]. Clark and FoxTree [11] find that different English fillers are used in different contexts and therefore they suggest to consider them as words.

Because fillers and filled pauses are frequent in spoken language, it is important to exploit their use and functions as well as their relation to gestures in order to include them in spoken language models which reflect the type of conversation and the language. The present study wants to contribute to these models by determining a) which are the most common fillers in a Danish corpus of first encounters, b) whether fillers co-occur with gestures and with which functions, and c) whether the most frequent fillers in Danish have conventionalized uses as in English and what these uses are.

In section 2, we discuss relevant related studies, then in section 3 we shortly describe the data and the methodology used for studying Danish fillers and filled pauses. In section 4, we present the analysis of fillers and filled pauses in the Danish corpus and , in section 5 we discuss the data. Finally, in section 6, we conclude and suggest future work.

2. Related studies

The functions of fillers in spoken language have been related to both interaction management and discourse planning. The various functions are not mutually exclusive and they are often related. Interaction management comprises feedback, that is feedback giving, also known as backchanneling, feedback eliciting [1, 12], and turn exchange regulation [2, 3, 11]. Turn exchange regulation comprises inter alia turn keeping and turn giving signals. Turn exchange signals mark also discourse planning processes. For example, a speaker can move her head away from the interlocutor signaling at the same time that she is planning her discourse and wants to keep the turn or the speaker can signal with a filled pause and gestures that she wants to give the floor if she has difficulties in completing the discourse.

Rochester [4] finds that filled pauses are more frequent when speakers face an option or have to express something challenging, while Reynolds and Paivio [13] report that students used pauses and filled pauses much more frequently when they had to define abstract objects than when they described concrete objects. Filled pauses can also mark the process of lexical retrieval [5] and researchers have noticed that the frequency of filled pauses is inverse proportional to the frequency of gestures [6, 7]. Esposito et al. [8] find that hand gestures co-occurring with English filled pauses involving the fillers *uh*, *um* and *ah* are often augmented holds, that is holds in which a little movement of the hand is noticed. They interpret the function of these holds as parallel to that of the speech pauses with which they co-occur. The speaker signals with the filled pause that she is planning new spoken content and marks with gestural holds that she is planning new gestures.

Language specific studies of fillers have focused on their type and their position in the utterance. For example, English

researchers have focused on the uses of the two fillers *uh* and *um*. More specifically, Shriberg [14] reports that vocal-nasal fillers are more frequent in the initial position of utterances in American English while vocal fillers occur most frequently when speakers have to find specific lexical items. Clark and FoxTree [11] propose to consider the English *uh* and *um* as words since speakers use them in a conventionalized way. The two researchers find that although the two fillers have many common uses, that is occur when speakers are looking for a word, planning the discourse, wanting to keep or give the floor, they have also a preferred or prototypical use. *Uhs* signal minor delays while *ums* signal major delays. Finally, Tottie [15] argues that *uh* and *um* can be used as discourse markers with a meaning similar to that of *well* and *you know*.

Swerts [16] analyzes the occurrences of filled pauses as markers of discourse boundaries in Dutch monologues, while De Leeuw [10] analyzes the realization of fillers in Dutch, English and German in order to determine their language specific characteristics. She finds that vocal-nasal fillers are predominant in English and German while vocal fillers are most common in Dutch. Vocalic-nasal fillers are only dominant in Dutch when they are surrounded by long pauses. English fillers are often preceded by a pause and followed by a lexical item, in De Leeuw's data, while in German and Dutch they are often surrounded by lexical items.

Possible effects of filled pauses on the listeners have also been investigated. For example, Fraundorf and Watson [17] prove that filled pauses have a positive effect on the listener's memory. Furthermore, different studies have determined that users perceive software agents to have more human-like behavior if they use fillers and therefore filled pauses have been included in the behavior of conversational software agents [18, 19, 20].

In a preceding study of pauses delimiting clause boundaries, and of the gestures which accompany them in the NOMCO corpus, we found that silent pauses and audible breath pauses are accompanied by head movements, facial expressions and body postures in 88% and 86% of their occurrences respectively, while filled pauses and pauses accompanied by other sounds are accompanied by the same gestures in only 77.5% and 70% of their occurrences respectively [21]. Furthermore, we found that the majority of clausal boundary pauses in the data were silent and breath pauses.

To our best knowledge, there are no previous general studies of fillers, filled pauses and the gestures which co-occur with them in Danish. However, it must be noted that the Danish fillers *hmm*, *øh* and *øhm* are included in a recent general language Danish lexicon *Den Danske Ordbog*[1]. In this lexicon, the three fillers are classified as interjections and are described as synonymous expressions of doubt. Furthermore, *øh* and *øhm* are analyzed as synonyms when used to fill in pauses while the speaker is thinking, and the filler *hmm* is defined as an interjection which expresses discontent, or a kind of disagreement or reservation with respect to the following word(s). In the following study of the functions of Danish fillers, filled pauses and co-occurring gestures, we will also investigate whether the lexicon definitions provided by the lexicon cover the uses of the fillers in the multimodal corpus of first encounters.

[1]Den Danske Ordbog is available on the internet at the address http://ordnet.dk/ddo/ordbog.

3. The data and method

The Danish NOMCO corpus consists of twelve multimodal annotated Danish first encounters which were collected and annotated under the Nordic NOMCO and the Danish VKK project. The NOMCO project's main aims were to create and analyze annotated comparable Nordic multimodal corpora, and first encounters were collected in more Nordic languages [22]. Furthermore, the conversations were annotated in all the corpora following a common theoretic framework [23], the so-called MUMIN annotation framework [24]. The Danish VKK project had the aim to analyze and model specific aspects of multimodal communication in Danish such as feedback and turn management [25, 26].

Six females and six males, aged 21-36 and native Danish speakers, were engaged in two encounters each, one with a female and one with a male. The participants talked freely about themselves, their studies and work while being audio and video recorded. Two microphones and three cameras were used and the encounters took place in a studio at the University of Copenhagen. Two snapshots from the data showing the three camera views are in Figure 1 and Figure 2.

Each encounter lasts between four and seven minutes, and the corpus has a duration of one hour. The annotations of the corpus comprise speech token transcription and shape and function descriptions of communicative co-speech gestures. In the speech transcriptions pauses are annotated as tokens and are annotated as a plus sign +. Furthermore, filled pauses, breath and other audible sounds accompanying pauses are also annotated.

The annotations of gestures are connected to speech tokens produced by either participant if the annotators found them to be semantically related. The gestures annotated are head movements, facial expressions and body postures [25]. For this study, we have added shape annotations of hand gestures co-occurring with fillers. The gestural functions considered in this study are feedback, self-feedback and turn management.

Table 1 shows the shape features of the gestures which are relevant to the present research while the function features of the gestures are in Table 2. The features describing the shape of gestures are coarse grained and only the most general shape features are used in this study. It must also be noted that information about gestural phases is not available.

Table 1: Shape features

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

The first two function features in Table 2 are related to feedback. The values of the attribute *FeedbackBasic* are assigned if feedback expresses Contact, Perception and Understanding (CPU) and if feedback only shows Contact or Contact and Perception but no Understanding (FeedbackOther) [1]. A positive feedback attribute is accompanied with the values of the *FeedbackDirection* attribute indicating whether feedback is given or

Figure 1: Two frontal snapshots from the corpus

Table 2: Function features

Attribute	Value
FeedbackBasic	CPU, FeedbackOther, None
FeedbackDirection	FeedbackGive, FeedbackElicit, None
TurnManagement	TurnTake,TurnHold, TurnAccept, TurnElicit, None

elicited.

The third function attribute, *TurnManagement* describes turn related behavior. The following four turn management values are relevant to the present study: a) TurnTake is assigned if the speaker signals that she wants to take a turn that wasn't offered; b) TurnHold: the speaker signals that she wishes to keep the turn; c) TurnAccept: the speaker signals that she is accepting a turn that is being offered; d) TurnElicit: the speaker signals that she is offering the turn to the interlocutor [26].

Inter-coder agreement were run on the data and resulted in Cohen's kappa scores in between 0.6 and 0.9 depending on the attributes. The transcriptions and annotations were made by one annotator, corrected by a second annotator and, in case of disagreement between the two main annotators, a third expert annotator took the final decision. We have used the final version of the data in this study. A more detailed description of the annotation procedure is in [25].

For the present study, we have identified all the fillers and filled pauses in the NOMCO corpus and we have extracted the co-occurring gestures with a perl script. Co-occurring gesture are defined as those gestures which temporally overlap with fillers or filled pauses. No limitation to the extension of the overlap were given. we have then extracted the duration of the fillers, and manually analyzed the context in which they occur, that is the speech tokens which precede and follow the fillers as well as the gestures which co-occur with them.

4. Analysis

There are 18,556 speech tokens (words, fillers and pauses) in the Danish first encounters, while there are 3,117 head movements, 1,448 facial expressions, 982 body postures and 566 hand gestures. The fillers in the Danish corpus are *øh*, *øhm*, *mm*, *årh*, *åh*, *hm/ehm*. Their frequency is in Table 3. Thus the most common filler is *øh*, *øhm* and *mm*.

Table 3: Filler types and their frequency

Filler	Occurrences
øh	375
mm/hmm	109
øhm	84
årh	9
åh	9
ehm	1
Total	587

Table 4 shows the fillers, their occurrences, their multimodal occurrences and the percentage of multimodal occurrences of fillers.

Table 4: Filler types and co-occurring gestures

Filler	Occurrences	Multimodal	%
øh	411	308	75
mm	113	92	81
øhm	91	70	77
årh	9	8	89
åh	9	9	100
ehm	1	1	100
Total	634	488	77

The number of the occurrences of the fillers in table 4 is higher than that in table 4 because when gestures are added to the speech tokens, some speech tokens are doubled. This is for example the case if two head movements co-occur with the same filler as indicated in figure 3.

Slightly over two-thirds of the occurrences of the fillers co-occur with gestures, and the number is the same as that of gestures co-occurring with filled paused [21]. In the rest of the study, we focus on the three most common fillers, that is the

Figure 2: A total view snapshot from the corpus

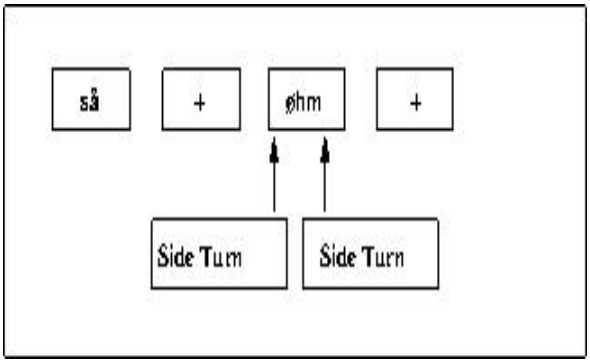


Figure 3: Two head movements co-occur with a filler

vocal *øh*, the nasal *mm* and the vocal-nasal *øhm*, on theirs uses and the types of gestures which co-occur with them.

Table 5 shows the percentage of *øhs*, *mms* and *øhms* which co-occur with head movements, facial expressions, body postures and hand gestures. In the table, it is not accounted for the fact that more gesture types can co-occur with the same filler occurrence.

Table 5: Filler types and co-occurring gestures

Filler	Occurrences	Head	Face	Body	Hand
øh	411	50%	18%	26%	11%
mm	113	68%	28%	19%	2%
øhm	91	55%	38%	23%	8%

The three most frequent fillers co-occur in most cases with head movements which are the most common body behavior. The filler *mm* co-occurs most frequently with head movements (half of its co-occurrences) while in 1/3 of the cases it co-occurs with facial expressions. The filler *øh* co-occurs frequently with head movements and body postures (50% and 26% respectively) and more seldom with facial expressions and hand gestures (18% and 11% of the occurrences respectively), while *øhm* co-occurs frequently with head movements, facial expressions and body postures (55%, 38% and 23% of the occurrences) and, less frequently that is in 8% of the occurrences with hand gestures.

In Table 6 are the mean and standard deviation of use of the three fillers. Table 6 shows that standard deviation for the use of

Table 6: Mean and standard deviation of fillers' use

Filler	Mean	Stand.dev.
mm	9.17	12.02
øh	31.64	27.84
øhm	7.75	3.49
all	48.42	25.2

fillers is quite large especially for the two fillers *øh* and *mm*, and some participants used more fillers than others. For analyzing the spoken contexts in which the fillers occur in the first encounters we have distinguished the following context categories:

- The filler occurs inside a phrase preceding a content word (adjective, adverb, noun, or verb). In this context the filler is often accompanied by a pause and signal lexical retrieval, e.g. A: *helt* + *øh* + *ubehøvlet* (completely + uh + boorish)
- The filler precedes a phrase or a feedback word such as *okay*, *ja* (yes) or *no*, e.g. + + *øhm jeg har også musiklinjen fra seminariet* (++ um, I have also music from the teacher school).In this context the filler is always accom-

panied by a longer pause and signals that the speaker is planning the discourse.

- The filler occurs in self-repairs and false starts. Also in these cases the filler is often accompanied by pauses, e.g. *læst + øh + hvor jeg læser*(studied + uh + where I study...)
- The filler (often a filled pause) occurs at the end of a turn, that is before the interlocutor takes the floor, E.g A: *det var sådan lidt +øh + (breath)* (it was such a little + uh + (breath), B: *(smack) + okay det junne du + det kan du godt lide* ((smack) + okay you did + you like it).
- The filler occurs alone as feedback marker or co-occurs with laughter, e.g. A: *er det på dansk?* (is it at Danish?) B: *mm* (yes).

In Table 7, we show the average duration in milliseconds of the three most frequently occurring fillers and of eventual pauses surrounding them. The percentage of occurrences of each filler in each spoken context is also given.

72% of the occurrences of the vocal filler *øh* are connected to speech planning in these data. More specifically in 50% of its occurrences, *øh* precedes a content word signaling lexical retrieval, while in 32% of the occurrences the filler precedes a phrase or a feedback word. *Øh* also occurs in self repairs (7% of the occurrences) and it precedes turn ends (9% of the cases), while only in 2% of its occurrences *øh* is used alone as feedback marker.

The functions of the gestures co-occurring with *øh* corresponds not surprisingly to the function of the filler indicated by the spoken context. In fact, the filler often co-occurs with gestures having a turn keeping function and, in fewer cases, a turn giving/yielding function. This finding confirms preceding studies that indicate that speakers signal with their body behavior that they want to keep the turn while searching for a word or while planning an utterance, i.a. [5, 27], or that speakers wish to give the floor if they have difficulties in completing the discourse [11]. *Øh* is only related to feedback gestures in 15% of its occurrences, and it co-occurs with self-feedback gestures, predominantly facial expressions, in 30 % of the cases.

The vocal-nasal *øhm* often precedes phrases or feedback words (70% of its occurrences). The majority of the phrases preceded by *øhm* are clauses and the filler often follows the conjunctions *og* (and) and *så* (so, therefore). *Øhm* occurs in the middle of a phrase in only 10% of the occurrences and in self repairs in 11% of the cases. Finally, it occurs at the end of a turn in 8% of its occurrences. The gestures that co-occur with *øhm*s are often feedback head movements (62% of the cases) and/or turn-management gestures (42% of the cases). In 30% of the occurrences gestures co-occurring with the filler *øhm* are facial expressions having a self-feedback gesture (own communication management).

The nasal filler *mm* occurs in most cases alone (65% of the occurrences) as feedback signal. It precedes a phrase or another feedback word in 35% of the occurrences. Similarly to *øhm*, in these cases it often follows the conjunctions *og* (and) or *så* (so, therefore) and precedes a clausal phrase.

Not surprisingly, the gestures which co-occur with *mm* are often related to feedback giving (backchanneling) (71% of the cases) and they are often nods. More rarely self-feedback and turn management gestures co-occur with *mm* (11% and 16% of the cases respectively).

A first analysis of half of the data indicates that holds in gestures often occur when fillers are related to lexical retrieval and discourse planning. There are no gestural holds when fillers are related to feedback giving and self-feedback.

5. Discussion

The main function of the filler *mm* is that of signaling feedback, alone or in connection with other feedback words. More seldom *mm* marks the start of a phrase. As feedback marker *mm* is often accompanied by feedback head movements, especially nods. The fact that this filler is mostly used as a feedback marker is also reflected by its average duration, which is shorter than that of the other two fillers. The use of the filler as feedback mark is in line with its analysis in the Danish dictionary *Den Danske Ordbog*, but in these data the filler is only connected to disagreement in four cases. In the majority of occurrences it simply signals feedback giving. This might be due to the type of interaction. Furthermore, different meanings of e.g. feedback words can only be identified when the multimodal context is available (audio and video), and the effect of intonation and gestures on the interpretation of the semantics of feedback words in Danish data has been proved previously [28].

The vocal-nasal filler *øhm* frequently occurs with pauses and precedes clauses or even utterances and therefore signals discourse planning processes. *Øhm* also occurs at the end of speech turns signaling that it has been interpreted by the interlocutor as a turn giving signal. *Øhm* also occurs with a certain frequency inside a phrase marking lexical retrieval (10% of the occurrences) or in self repairs (11% of occurrences), while it only seldom occurs alone as feedback marker (1% of the occurrences). *Øhm* lasts longer than the other two fillers (0.48 milliseconds in average) and this finding is in line with what has been noted about English data: English filled pauses that mark larger syntactic units have a longer duration that filled pauses that signal lexical retrieval inter alia [11, 15]. It is not strange that the gestures which co-occur with *øhm* have feedback and turn management functions since feedback and turn management signals often occur at utterance or clausal boundaries.

The most common filler *øh*, which has an average duration between that of the two other fillers (0.4 milliseconds) most often precedes a content word, signaling lexical retrieval and, less frequently, precedes a phrase in these data. Other uses of *øh* mark self repairs or turn end. Only seldom *øh* is used as feedback marker (2% of its occurrences). The most common function of the gestures co-occurring with *øh* is that of self-feedback. This is not surprising since speakers often produce self-feedback gestures in self repairs or at the end of their spoken contributions.

Even though all fillers in the Danish data occur as signals in interaction management and/or discourse planning contexts as it was the case for fillers in other languages [10, 11], each filler has some more common or prototypical uses, as it also is the case for English fillers [11]. Our data indicates that the use of fillers variates from one participant to the other, and some participants use more frequently one or two fillers. The standard deviation was particularly high for the two fillers *øh* and *mm*, while it was lower for the filler *øhm*. The reason for this variation should be investigated in the future.

Thus, the analysis of the first encounters confirms overall the uses of the three fillers described in the Danish lexicon, but it also indicates that even though *øh* and *øhm* can occur in the same contexts and can be used as synonyms, they have different preferred/prototypical uses. Moreover, the three fillers have even more uses than those described in the Danish lexicon. Interestingly, our data indicate that the prototypical uses of the

Table 7: Fillers, duration and spoken contexts

Filler	Duration	Inside phrase	Before phrase	Self repair	End turn	Alone
øh	0.4	50%	32%	7%	9%	2%
øhm	0.48	10%	70%	11%	8%	1%
mm	0.3	0	35%	0	0	65%

two Danish fillers *øh* and *øhm* are the same as those of the English fillers *uh* and *um* [11, 15] that is *øh* and *uh* often precede a content word and signal lexical retrieval while *øhm* and *um* precede a clausal phrase or an utterance signaling planning of a larger discourse part and having a function similar to discourse markers [15]. This is also in line with studies on silent pauses. For example Tøndering [29] finds that silent pauses preceding subordinated phrases are shorter than those between independent phrases in a Danish spoken corpus, the DANPASS corpus.

The fact that gesture holds were only found when the gestures co-occurred with filled pauses confirms the study by Esposito et al. [8] which found that augmented hand gestural holds co-occurred with speech pauses.

6. Conclusion and future work

In the paper, we presented a study of fillers, filled pauses and co-occurring gestures in the Danish NOMCO corpus of first encounters. In these data, the majority of the fillers are accompanied by gestures and the gestures reinforce the filler's function.

The Danish fillers have the same functions as fillers in other languages, that is they have functions related to feedback and/or turn management, or they signal discourse planning processes, hereunder lexical retrieval. The various functions are not mutually exclusive.

The analysis of the Danish data shows that each filler type has a predominant use, even though the most common fillers are often used synonymously, that is they can also occur in the same contexts. This finding confirms for Danish what has been also found to be the case for fillers in other German languages and especially in English.

More specifically, we found that the nasal filler *mm* is often used as a feedback giving marker and it is nearly always accompanied by feedback head movements. Moreover, in this corpus, it mostly indicates positive feedback and co-occurs with nods. This is not surprising since in first encounters participants are kind and try to give the interlocutor a positive impression [30].

The vocal *øh* often signals lexical retrieval, but it is also used in other contexts. The gestures which accompany this filler have mainly turn management functions or signal self-feedback. *Øh* has the same prototypical use as the English vocal filler *uh*.

The vocal-nasal filler *øhm* often precedes a clausal phrase and marks discourse planning. It is often accompanied by gestures having an interaction management function. The prototypical function of *øhm*is the same as that of the English nasal-vocal filler *um*. As in English, filled pauses occurring at the boundaries of larger discourse units have longer duration than filled pauses preceding lexical entries.

In the future, we will analyze fillers in more types of spoken data, including monologues and dialogs involving more than two participants as well as in conversations between interlocutors who know each other in advance. Individual differences in the use of fillers and filled pauses should also be investigated and the uses and occurrences of fillers and filled pauses in the Danish first encounters should be compared with the uses of fillers in the other Nordic first encounters corpora.

7. Acknowledgments

I would like to thank my colleague Patrizia Paggio, the Danish NOMCO and VKK annotators Anette Luff Studsgård, Sara Andersen, and Bjørn Wessel-Tolvig. Special thanks also go to the NOMCO project's Nordic partners, Elisabeth Ahlsén, Jens Allwood and Kristiina Jokinen.

8. References

[1] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[2] H. Maclay and C. E. Osgood, "Hesitation phenomena in spontaneous English speech," *Word*, vol. 15, pp. 19–44, 1959.

[3] S. Duncan and D. Fiske, *Face-to-face interaction*. Hillsdale, NJ: Erlbaum, 1977.

[4] S. R. Rochester, "The significance of pauses in spontaneous speech," *Journal of Psycholinguistic Research*, vol. 2, pp. 51–81, 1973.

[5] R. Krauss, Y. Chen, and R. F. Gottesman, "Lexical gestures and lexical access: a process model," in *Language and gesture*, D. McNeill, Ed. Cambridge University Press, 2000, pp. 261–283.

[6] N. Christenfeld, S. Schachter, and F. Bilous, "Filled pauses and gestures: It's not coincidence," *Journal of Psycholinguistic Research*, vol. 20, no. 1, pp. 1–10, 1991.

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

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

[9] D. McNeill, *The Conceptual Basis of Language*. Routledge Library Editions: Linguistics, 2014.

[10] E. de Leeuw, "Hesitation Markers in English, German, and Dutch," *Journal of Germanic Linguistics*, vol. 19, pp. 85–114, 6 2007.

[11] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–11, 2002.

[12] J. Allwood, "Dialog Coding - Function and Grammar: Gteborg Coding Schemas," *Gothenburg Papers in Theoretical Linguistics, University of Gteborg, Dept of Linguistics*, vol. 85, pp. 1–67, 2001.

[13] A. Reynolds and A. Paivio, "Cognitive and emotional determinants of speech," *Canadian Journal of Psychology*, vol. 22, pp. 164–175, 1968.

[14] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, Berkeley, 1994.

[15] G. Tottie, "Uh and um in British and American English: Are they words? Evidence from co-occurrence with pauses," in *Linguistic Variation: Confronting Fact and Theory*, N. Dion, A. Lapierre, and R. T. Cacoullos, Eds. New York: Routledge, 2014, pp. 38–54.

[16] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, pp. 485–496, 1998.

[17] S. Fraundorf and D. Watson, "The disfluent discourse: Effects of filled pauses on recall," *Journal of memory and language*, vol. 65, no. 2, pp. 161–175, 2011.

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

[19] D. Traum and J. Rickel, "Embodied agents for multi-party dialogue in immersive virtual worlds," in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2*, ser. AAMAS '02. New York, NY, USA: ACM, 2002, pp. 766–773.

[20] L. Pfeifer and T. Bickmore, "Should Agents Speak Like, um, Humans? The Use of Conversational Fillers by Virtual Agents," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjálmsson, Eds. Springer Berlin Heidelberg, 2009, vol. 5773, pp. 460–466.

[21] C. Navarretta, "Pauses delimiting semantic boundaries," in *Proceedings of the 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2015)*, IEEE, Ed., Giör, Hungary, October 2015, pp. 533–538.

[22] P. Paggio, E. Ahlsén, J. Allwood, K. Jokinen, and C. Navarretta, "The NOMCO multimodal Nordic resource - goals and characteristics," in *Proceedings of LREC 2010*, Malta, May 17-23 2010, pp. 2968–2973.

[23] C. Navarretta, E. Ahlsén, J. Allwood, K. Jokinen, and P. Paggio, "Feedback in Nordic First-Encounters: a Comparative Study," in *Proceedings of LREC 2012*, Istanbul Turkey, May 2012, pp. 2494–2499.

[24] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing," *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation*, vol. 41, no. 3–4, pp. 273–287, 2007.

[25] P. Paggio and C. Navarretta, "Head Movements, Facial Expressions and Feedback in Danish First Encounters Interactions: A Culture-Specific Analysis," in *Universal Access in Human-Computer Interaction- Users Diversity. 6th International Conference. UAHCI 2011, Held as Part of HCI International 2011*, ser. LNCS, C. Stephanidis, Ed., no. 6766. Orlando Florida: Springer Verlag, 2011, pp. 583–690.

[26] C. Navarretta and P. Paggio, "Classifying Multimodal Turn Management in Danish Dyadic First Encounters," in *Proceedings of the 19th Nordic Conference of Computational Linguistics (Nodalida 2013)*. Oslo, Norway: NEALT, May 2013, pp. 133–146.

[27] A. Kendon, *Gesture - Visible Action as Utterance* . New York: Cambridge University Press, 2004.

[28] C. Navarretta and P. Paggio, "Classification of feedback expressions in multimodal data," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Upssala, Sweden, Juli 2010, pp. 318–324.

[29] J. Tøndering, "Prosodiske fraser og syntaktisk struktur i spontan tale," *NyS - Nydanske Sprogstudier*, vol. 39, pp. 166–198, 2010.

[30] C. Navarretta, "Annotating and analyzing emotions in a corpus of first encounters," in *Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunications*, IEEE, Ed., Kosice, Slovakia, December 2012, pp. 433–438.

Starting a Conversation with Strangers in Virtual Reykjavik: Explicit Announcement of Presence

Stefán Ólafsson[1,3], *Branislav Bédi*[2], *Hafdís Erla Helgadóttir*[1], *Birna Arnbjörnsdóttir*[2], *Hannes Högni Vilhjálmsson*[1]

[1]CADIA, Reykjavik University
[2]University of Iceland
[3]Relational Agents Group, Northeastern University

stefanola13@ru.is, brb19@hi.is, hafdis13@ru.is, birnaarn@hi.is, hannes@ru.is

Abstract

Virtual Reykjavik is an Icelandic language and culture training application for foreigners learning Icelandic. In this video game-like environment, the user is asked to solve given tasks in the game and in order to complete them he/she must interact with the characters, e.g. by conversing with them on context-specific topics. To make this a reality, a model for how natural conversations start in a specific situations has been developed, based on data from that same situation in real life: a stranger asking another stranger for directions to a particular place in downtown Reykjavik. This involved defining a multimodal annotation scheme, outlining the communicative functions and behaviors associated with them. However, current annotation schemes lacked the appropriate function for this specific case, which lead us to finding and proposing an appropriate communicative function – the Explicit Announcement of Presence. A study was conducted to explore and better understand how conversation is initiated in first encounters between people who do not know each other. Human-to-human conversations were analyzed for the purpose of modelling a realistic conversation between human users and virtual agents. Results from the study have lead to the inclusion of the communicative function in the human-to-agent conversation system. By playing the game the learners will be exposed to situations that they may encounter in real life, and therefore the interaction is based on real life data, rather than textbook examples. We believe that this application will help bridge the gap from the class room to the real world, preparing learners to initiate conversations with real Icelandic speakers.

Index Terms: Explicit Announcement of Presence, communicative function, human-agent interaction, embodied conversational agent, multimodal communication, natural language, social behavior

1. Introduction

The Icelandic language and culture training application *Virtual Reykjavik* is an on-line computer game environment supporting game-based learning [1], task-based learning [2] and a communicative approach [3, 1] to teach Icelandic as a foreign language (adult learners living outside of Iceland) or second language (adult learners living in Iceland). Learners (users from now on) can gain particular linguistic and cultural skills by engaging in interactive exercises and are then able to use that knowledge in conversations with real people in the natural setting of the target language [4]. The exercises entail practicing saying words and phrases in simple conversations with *Embodied Conversational Agents* (ECAs), which are defined as a computer interface represented by a humanoid body that is specifically conversational, exhibiting and recognizing the behavior involved during human face-to-face conversation [5]. In these interactions, the users find themselves in various situations, such as encountering a stranger, starting a conversation, and asking him/her for directions (see Figure 2).

Figure 1: *A screenshot of an ECA in Virtual Reykjavik. A yellow arrow appears overhead when the user has targeted the ECA and the mouse may be clicked to activate speech recognition. The green, yellow, and red lights in the upper right corner indicate the user's changing role in the interaction, i.e. listener or speaker.*

In Virtual Reykjavik, users interact with different ECAs (male or female) in the following ways: by approaching an agent until it acknowledges the user's presence, using the mouse to signal which direction or which ECA the user is looking at, clicking the mouse to trigger an action, such as speaking, and by talking into the microphone through which the ECA gets the speech input from the user (Figure 1). Similar to the Tactical Language and Culture Training System [4], Virtual Reykjavik also relies on natural spoken language when interacting with game characters, i.e. ECAs, via automatic speech recognition. For Icelandic, the current version of our system uses the

Figure 2: *The first image (from the left) shows the ECA's reaction to the user performing behavior associated with the Explicit Announcement of Presence (EAP), the second shows the agent's reaction to a question, the third shows the agent answering the question, and the last image shows the reaction to the user saying thank you.*

Google speech recognition service[1], because it is currently the only readily available software for the language [6]. In the first learner scenario or level of the game, users have to fulfill three tasks: 1) to get an ECA's attention and start a conversation, 2) to ask the ECA for directions, and 3) to say goodbye to the ECA (see Figure 2). In this article, we discuss the finding and implementation of an appropriate communicative function associated with getting a stranger's attention and the multimodal behavior associated with the acknowledgement of it.

2. Motivation

In situations where participants know each other, a greeting phase often fulfills the function of noticing and acknowledging one's presence and initiating a conversation [7]. During exploratory data collection for the Virtual Reykjavik project, we found that when strangers approach one another and start a conversation, something other than a greeting occurs. We were motivated to take a closer look at what behaviors native speakers of Icelandic exhibit and what communicative functions they carry out, in order to provide the users of Virtual Reykjavik with an accurate portrayal of conversations in the language.

We recorded and annotated naturally occurring human-to-human conversations in order to faithfully emulate conversational behavior in Virtual Reykjavik. During the annotation process, we felt that the communicative function being conveyed at the very beginning of each encounter was not to be found in current standard multimodal annotation schemes, such as the MUMIN coding scheme [8], SmartKom multimodal corpus [9] or the HuComTech multimodal corpus annotation scheme [10].

In the theoretical exploration of this topic, we came upon the *Explicit Announcement of Presence* (EAP) [7], which we used for building our hypothesis: when strangers meet during first encounters in situations when they ask for directions, the Explicit Announcement of Presence is the communicative function underlying the behavior at a start of a conversation.

3. Modelling Approach

In order to design a realistic conversational structure for our context-specific situation using the Icelandic language, we needed to address two general problems. First, we needed to define the appropriate communicative functions and behaviors that would best fit our context, i.e. how a stranger (non-native speaker) approaches another stranger (native speaker) in downtown Reykjavik, and how the approached person acknowledges it. We then needed to implement the function involved in getting someone's attention and the natural multimodal behavior associated with acknowledging it in the virtual characters. A traditional greeting-phase, often used as initial learning scenarios in textbooks, would not apply in the situation we picked, because traditional greetings are primarily used amongst persons who know each other. On that account we needed to come up with something new.

We worked towards realizing a conversational structure that would maintain presence and authenticity, with the aim of giving the user a feeling of a natural conversation akin to what we observed in our field study. The approach we took was inspired by Clark's (1996) *conversation sections*, which are purpose-specific segments of a conversation that arise during the course of face-to-face interaction between humans. On a very high level, these sections include the *entry*, *body*, and the *exit* of the conversation [11]. However, thinking of the body of a conversation as one single purpose-specific section is rather vague. Thus, for our purposes, we tried to identify portions of the conversations in our data as being potential conversation sections, portions where the participants are bound to an identifiable purpose.

An example of such a segment in a real-life context is during the task of asking a question and receiving an answer. The initiating participant has a purpose, i.e. to gain some knowledge from the other party, and in the process he/she alters the intent of the other. In other words, the initiator influences other participant's intentions and together they become involved in this purpose-specific segment, or conversation section.

The emergence of a conversation section at any given time during a conversation is governed by multiple factors, such as the relations between the participants, their intention, and personality. Moreover, all of these factors affect what functions and behaviors are involved in the context of the particular conversation section at hand. For instance, in an informal setting where participants know each other, a greeting would sound and look different to one in a formal setting where participants do not know each other. We defined the EAP as the appropriate function for initiating conversation in the following setting: (1) participants who are strangers; (2) a non-native speaker approaches a native speaker in an informal setting, which is downtown Reykjavik, and asks for directions to a particular place.

[1] http://goo.gl/eSRnbv

Track Type	Function Category	Type
Interactional	Initiate	react, recognize, salute-distant, salute-close, initiate
	Turn-taking	take, give, keep, request, accept
	Speech-act	eap*, inform, ask, request
	Grounding	request-ack, ack, repair, cancel

Table 1: These are the in interactional function categories from the original FML proposal [12] for use in the Virtual Reykjavik system (alteration marked with *).

4. EAP Study

4.1. Method

In order to better understand the use of EAPs for initiating conversations, both in terms of frequency of use and how they are manifested in behavior, we conducted a small qualitative study. Natural language data from conversations of first encounters was collected in the form of video recordings. The focus was on approaching a stranger and starting a conversation. Two volunteer actors, both female non-native speakers of Icelandic, were hired to approach Icelanders and ask for directions to a particular place in downtown Reykjavik. The first human-to-human conversations we recorded were done by walking up to people and stating our purpose beforehand. This made it impossible to capture the initial moments of naturally occurring contact. We therefore changed our method to stating the purpose of our research to people after the conversation. The actors received only one instruction: to ask people for directions. Without further telling the actors what to do, they started naturally approaching people and announcing their presence. Consent from participants was recorded on camera at the end of each recording and participants could ask to withdraw from the study and their recording would be deleted on the spot.

The actors were asked to conduct themselves as normally as possible. The effect of them being non-native speakers is negligible in these circumstances, since in all cases they performed the appropriate utterances and had clear pronunciation. The selective sampling method [13] was applied here in order to address the right group of people and ensure the authenticity of the collected data. Only male and female native speakers of Icelandic aged between 18-70+ were considered. The study was anonymous and concession was received from all participants.

The video recordings were annotated using a multimodal annotation scheme for Virtual Reykjavik compiled from various other research (see Figure 3), both in terms of the communicative behavior present in the dialogue and the underlying intent or function of those behaviors [14, 15, 16, 17, 18, 19, 12, 20, 21, 22, 23]. This follows the distinction between function and behavior made in the SAIBA framework for multimodal generation of communicative behavior for ECAs, as manifested in the Behavior Markup Language (BML) [18] and Function Markup Language (FML) [12]. Our current work contributes to existing work on FML by introducing the EAP as a type of a communicative function. In our observations, a verbal behavior typically follows the EAP. We therefore categorize the EAP as a type of *speech-act*, which is a communicative function category that includes multiple types (*ask*, *inform*, etc.), adding to the current FML standard (see Table 1).

4.2. Results

We analyzed 44 videos that included first encounters between native speakers and non-native speakers of Icelandic asking for directions to a specific place in downtown Reykjavik. The focus was on the first part of the dialogue, i.e. approaching a person and initiating a conversation. The data shows that in 33 videos (75% cases) pedestrians passing through (non-natives) announce their presence verbally to other unknown pedestrians (natives), in 10 videos (23% cases) both notice each other before an announcing phase has a chance to happen, and in 1 video (2% cases) the phase was described as "other" because it could not be identified (see Figure 4).

Results show that in most of the cases approaching pedestrians announce their presence verbally in order to cause attention from the approached pedestrian to initiate a conversation about getting directions to a particular place. It became clear that a particular communicative function – the EAP – was primarily being conveyed verbally here by the non-natives when approaching the natives.

The most frequent EAPs in our data have the folowing form: 1) phrases: *fyrirgefðu* [pardon me], *afsakið* [excuse me], 2) greetings: *góðan daginn* [good day] with definite article / *góðan dag* [good day] without definite article / *hæ* [hi], or 3) directly asking the question: *Veistu hvar X er?* [Do you know where X is?]. In our study, explicit nonverbal EAPs were not

Multimodal Data for Modeling Agent's Response to EAP

SPEECH	--
HEAD	Central, directed at speaker / user
FOREHEAD	Slightly crumpled
EYE BROWS	Slightly raised, slightly drawn together
GAZE	Direct, open eyes
MOUTH	Slightly open
HANDS	Beside the body, no movement
BODY POSTURE	Directed at the speaker / user, aligned with the torso
DISTANCE	Close to the speaker / user
TORSO	Directed at the speaker / user, aligned with the whole body
LEGS / FEET	Finish movement (walking), legs slightly apart when standing still

Figure 3: *An example description of behavior using an annotation scheme developed for Virtual Reykjavik.*

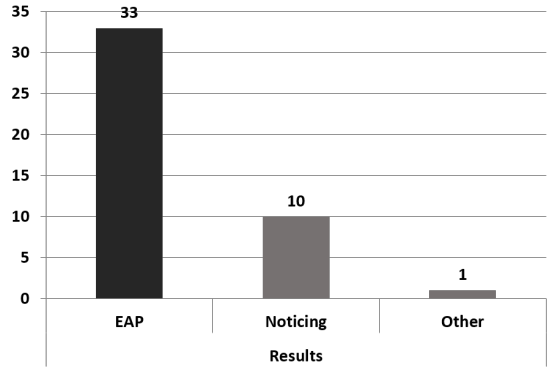


Figure 4: *Most people starting an interaction with a stranger in the street (33 out of 44 videos) show behavior that carries out an EAP function.*

found, except when looks of approaching participants accidentally met. But this has been categorized in FML as *noticing*, because it involves a stranger (actor) gazing at another stranger (native speaker), who responds by gazing back and awaiting some kind of a response from the gazing actor. All of this takes place in a fraction of a second.

After the EAP is performed by the non-native stranger, the person being approached generally realizes that someone wants to speak to him/her and looks back at that person. Detailed analysis of the multimodal behavior, exhibited by the native speakers as a result of the EAP by the non-native speakers, was performed using representative subjects: one male and one female native speaker, both around 50 years of age. The data was annotated in Elan [24], and the Multimodal Annotation Scheme for Virtual Reykjavik was used as a reference. Results are listed in Table 2 and Table 3.

These results lead to the incorporation of the EAP communicative function into the Virtual Reykjavik ECAs, including the realization of plausible EAP related behavior.

5. Implementation

The EAP is just one of many communicative functions, among others such as *turn-taking* and *grounding*, that precede a set of one or more behaviors. The communicative plans of an ECA are manifested in communicative functions, i.e. pieces of intent that have a communicative purpose. Based on these functions, the system then plans out which behaviors carry them out.

While the functions themselves are unseen, the behavior is their visible result. If a person wants to approach another person who is a stranger, their brain plans for an EAP and when the time comes it tells the body to perform the behaviors associated with that function, e.g. to look at the other participant and say *afsakið* [excuse me].

Implementing the EAP function within the conversational system architecture of Virtual Reykjavik lets the users interact with the ECAs in a more realistic way, and the ECAs get to exhibit realistic behavior in response to it. This behavior also relates directly to one of the important tasks that users have to perform in order to fulfill the game objectives, i.e. engage with a stranger in the street and ask him/her for directions.

Our implementation involved the use of conversation sections. Within our system, these sections are called blocks and are the objects that contain methods for producing communicative functions that underlie the behavior in various situations [25]. Knowledge regarding which behaviors and functions are appropriate for each situation was gathered from the annotated video data (see Figure 3 and Tables 2 & 3).

As mentioned above, the *entry* is the first purpose-specific segment of the conversation and a block in our system that corresponds to that is the *Approach* block (see Figure 5). This block, based on the observed data, necessarily includes the EAP in order for the stranger to initiate a conversation with an ECA [25].

The block element structure allows the Virtual Reykjavik conversation system to procedurally select what comes next in the conversation. The blocks provide a context for the communicative functions at any given moment and align speech with other modalities, in our case the conversational behavior of the ECAs. We had to design a system architecture that allows the agent to make a decision as to what should happen next in the conversation, based on dialog history, personality, and what events have unfolded in the interaction with the user at any given time.

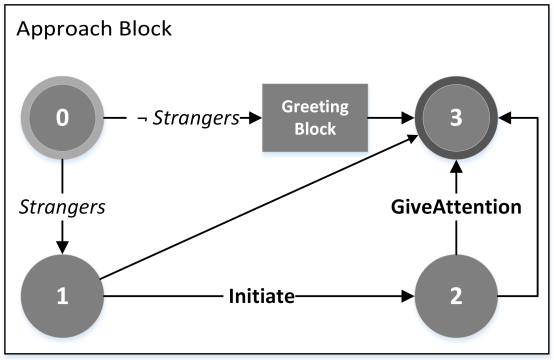


Figure 5: *The 'approach' block's state machine propels the conversation using methods (Initiate and GiveAttention, shown in bold) that generate discourse functions relative to the agents' intent. The initial state checks for relations and moves to either a greeting phase or a 'stranger specific' initiation of conversation. States (1) and (2) allow for 'inaction', resulting in the approach coming to an abrupt end in the final state (3).* [25]

This system allows the agents to either move to the next state within a particular block or, if the current block is finished, select which block of conversation they want to push next to the floor of interaction.

5.1. Initiating a Conversation

The following provides a more in-depth description of what transpires when the user approaches an agent in the first Virtual Reykjavik game scenario. When the user starts the program he/she embodies an avatar that is structurally very similar to the other ECAs in the scene, i.e. they both have perception systems that function in the same way and they perform behaviors in the same manner. The only difference is that the human user is in control of his/her avatar's head movement, where he/she walks, and does certain actions with the keyboard and the mouse.

When the user (player) moves his/her avatar closer to the agent, their respective perception systems perceive the other and their reasoning faculties check their intentions in order to decide whether to act on them (see Figure 6). While the agent has no interest in initiating a conversation, the player's intent for getting the information is made known by clicking the mouse when a yellow arrow appears above the agent's head, as in Figure 1. This prompts the speech recognition software to allow the player to speak and when he/she is done speaking the speech recognition automatically stops listening. The input is stored for further analysis and may have an impact on which block will be selected next.

Following this action, the program instantiates a *discourse manager* and a *floor* of interaction is created with the agent and player as participants. The discourse manager asks the floor to execute the next action in the current block; however, in this case, it finds that no current block is available. Therefore, the first block is established by looking at both participants' intentions and personality parameters and in this case an Approach block is selected.

The player and agent's relationship is checked in the first state of the Approach, and here they are found to be strangers. The participants then progress to the next state where the player's avatar creates a bundle of communicative functions

Nonverbal Reaction of B to A's EAP		**Description**
Head		central, directed at A
Face	Forehead	crumpled
	Eyebrows	slightly raised & slightly drawn together
	Eyes	open & directed at A
	Mouth	slightly open
Torso		slightly turned away from the A
Hands		beside the body, holding hands, no movement
Body posture		aligned with the torso = slightly turned away from the A due to A's interfering form the side
Position		close to the A

Table 2: *Sample nonverbal reaction of female native speaker of Icelandic (B) to the EAP of approaching female non-native speaker (A)*

Nonverbal Reaction of B to A's EAP		**Description**
Head		central, directed at A
Face	Forehead	crumpled
	Eyebrows	slightly raised & slightly drawn together
	Eyes	open & directed at A
	Mouth	slightly open
Torso		directed at the A
Hands		beside the body, holding hands, no movement
Body posture		aligned with the torso = directed at the A due to A's interfering,directly in the pathway of the pedestrian
Position		close to the A

Table 3: *Sample nonverbal reaction of male native speaker of Icelandic (B) to the EAP of approaching female non-native speaker (A)*

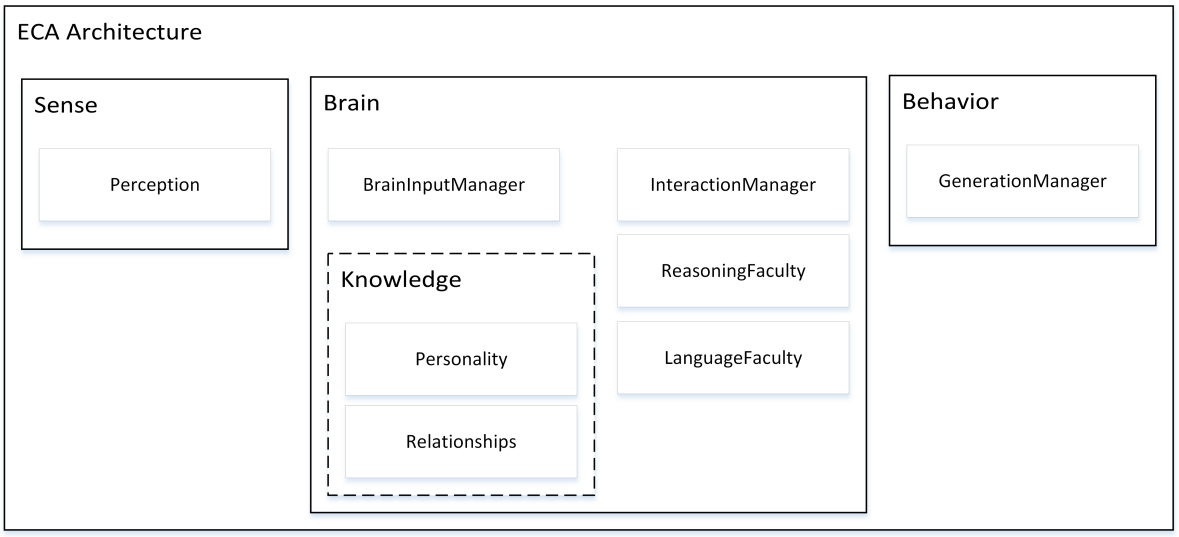


Figure 6: *The Virtual Reykjavik ECA architecture. The Perception component acts as the agent's sensor and communicates with the BrainInputManager. The Brain's components work together with the discourse system in making communicative functions to be sent to the GenerationManager for behavior realization.* [25]

called an FML document, which crucially includes the EAP. Following this, the other participant adds whatever functions he/she feels is necessary to the FML document, and finally the document is sent to each of the participants' behavior generation modules for processing. At this point, each ECA turns their respective FML information into BML (see section 4.1) and executes the relevant animations on the character.

Now the Approach block has reached an end state. The next time the floor calls for the the next action to be executed, a new block must be selected. It is not predetermined which block that will be. It is important to note that if the two participants had known each other, they would have gone down a different path within the Approach block, entered into a greeting phase, and the EAP would never have happened.

6. Pilot User Study

A pilot user study has been conducted where six non-native speakers of Icelandic (four female, two male) played the first scenario in Virtual Reykjavik. Five of the subjects were beginners and one at intermediate level in Icelandic. In the context of the first task - to get someone's attention - the following words, although correct Icelandic, were used incorrectly in this context: *sæll* [how are you[2]] said to a male, *sæl* [how are you] said to a female, *blessaður* [how are you] said to a male person, *blessuð* [how are you] said to a female person, *halló* [hello], and *hæ* [hi]. The agents did not respond adequately because they were not designed for such greetings that are usually used among friends, acquaintances, and persons who know each other. These preliminary results indicate that the students in this sample were taught how to greet, but perhaps not how to approach a stranger on the streets as people do in real life. Further experimentation is needed to validate these findings.

The results also revealed that each user also used one of the three types of the EAP's verbal forms (see section 4.2) to announce his/her presence when approaching an ECA. In some cases, however, the users only approached the ECA and waited until it notices them. The proximity to the agent served the purpose of getting noticed. As it was a pilot study, in preparation for further testing, recordings of the computer screen were not made and therefore precise information on the proxemics was not retained, but will be included in the future.

7. Conclusion and Future Work

When teaching foreigners a new language, like Icelandic, it is imperative that they get lessons that reflect what happens in actual conversation. When analyzing situations where a stranger approaches another stranger, it became clear that the classic greeting phase [7] was missing. In Icelandic language lessons, foreigners are taught how to greet others [26]; however, this is not what we observed native speakers doing when non-native speakers, who were strangers, started conversations with them.

We observed that the EAP was the communicative function that most frequently occurred in situations where a stranger sought to initiate a conversation with another stranger for the purposes of asking for directions. This prompted the inclusion of such a function within the discourse models that arise during human-to-agent interaction. A model was implemented whereby the user EAP was the catalyst for conversation. Approaching an agent and clicking the mouse calls for an EAP, which prompts the user to speak and the conversation begins.

Early pilot tests have revealed that users may use inappropriate vocabulary when approaching native speakers in the simulated natural environment Virtual Reykjavik. This kind of vocabulary included greetings used among people who know each other and therefore not suitable for the EAP. Whatever the cause, future versions of the ECAs need to be aware of this tendency and be able to give the students constructive feedback. On the basis of our study, the EAP can potentially be generalized to other languages, because it seems to be a natural way how strangers approach other strangers in situations when they want to ask a question, e.g. directions to a particular place.

Acknowledgements

This work was made possible thanks to the Icelandic Research Fund (RANNÍS) and the collaborative spirit of researchers at the University of Iceland and the Center for Analysis and Design of Intelligent Agents (CADIA) at Reykjavik University.

[2]There are not direct translations for these greetings, but they are forms not uttered between strangers

8. References

[1] B. Meyer, "Designing serious games for foreign language education in a global perspective," *Support for Learning*, vol. 1, pp. 715–719, 2009.

[2] R. Ellis, *Task-based language learning and teaching*. Oxford University Press, 2003.

[3] W. Littlewood, *Communicative language teaching: An introduction*. Cambridge University Press, 1981.

[4] L. W. Johnson, H. Vilhjalmsson, and S. Marsella, "Serious games for language learning: How much game, how much AI?" *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 2005.

[5] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "Embodied conversational agents." Cambridge, MA, USA: MIT Press, 2000, ch. Human Conversation As a System Framework: Designing Embodied Conversational Agents, pp. 29–63.

[6] J. Guðnason, O. Kjartansson, J. Jóhannsson, E. Carstensdóttir, H. Vilhjálmsson, H. Loftsson, S. Helgadóttir, K. Jóhannsdóttir, and E. Rögnvaldsson, "Almannaromur: An open icelandic speech corpus," in *Proceedings of the Third International Workshop on Spoken Language Technologies for Under-resourced languages (SLTU 2012)*, 2012.

[7] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge: Cambridge University Press, 1990.

[8] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena," *Language Resources and Evaluation*, vol. 41, no. 3/4, pp. 273–287, 2007.

[9] F. Schiel, S. Steininger, and U. Türk, "The smartkom multimodal corpus at bas." in *LREC*, 2002.

[10] K. Pápay, S. Szeghalmy, and I. Szekrényes, "Hucomtech multimodal corpus annotation," *Argumentum*, vol. 7, pp. 330–347, 2011.

[11] H. H. Clark, *Using Language*. Cambridge: Cambridge University Press, 1996.

[12] A. Cafaro, H. H. Vilhjálmsson, T. Bickmore, D. Heylen, and C. Pelachaud, "Representing communicative functions in saiba with a unified function markup language," in *Intelligent Virtual Agents*. Springer, 2014, pp. 81–94.

[13] I. T. Coyne, "Sampling in qualitative research. purposeful and theoretical sampling; merging or clear boundaries?" *Journal of advanced nursing*, vol. 26, no. 3, pp. 623–630, 1997.

[14] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, C. Soria, and D. Traum, "Language resource management—semantic annotation framework—part 2: Dialogue acts," *International Organization*, 2010.

[15] I. Zwitserlood, A. Ozyurek, and P. M. Perniss, "Annotation of sign and gesture cross-linguistically," in *6th International Conference on Language Resources and Evaluation (LREC 2008)/3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. ELDA, 2008, pp. 185–190.

[16] H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud *et al.*, "The behavior markup language: Recent developments and challenges," in *Intelligent virtual agents*. Springer, 2007, pp. 99–111.

[17] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari, "Multimodal human discourse: gesture and speech," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 9, no. 3, pp. 171–193, 2002.

[18] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *Intelligent virtual agents*. Springer, 2006, pp. 205–217.

[19] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson, "The next step towards a function markup language," in *Intelligent Virtual Agents*. Springer, 2008, pp. 270–280.

[20] A. Cafaro, "First impressions in human-agent virtual encounters," 2014.

[21] J. Allwood, L. Cerrato, L. Dybkjaer, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin multimodal coding scheme," *NorFA yearbook*, vol. 2005, pp. 129–157, 2005.

[22] S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin, "Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces," in *HCI International*, 2005.

[23] Á. Abuczki and E. B. Ghazaleh, "An overview of multimodal corpora, annotation tools and schemes," *Argumentum*, vol. 9, pp. 86–98, 2013.

[24] H. Sloetjes and P. Wittenburg, "Annotation by category: Elan and iso dcr." in *LREC*, 2008.

[25] S. Ólafsson, "When strangers meet: Collective construction of procedural conversation in embodied conversational agents," *Master's thesis, The University of Iceland, Reykjavik, Iceland*, 2015.

[26] "Icelandic online," http://icelandiconline.is/index.html, accessed: 2016-15-01.

Coordination of head movements and speech in first encounter dialogues

Patrizia Paggio

University of Copenhagen and University of Malta
paggio@hum.ku.dk, patrizia.paggio@um.edu.mt

Abstract

This paper presents an analysis of the temporal alignment between head movements and associated speech segments in the NOMCO corpus of first encounter dialogues [1]. Our results show that head movements tend to start slightly before the onset of the corresponding speech sequence and to end slightly after, but also that there are delays in both directions in the range of -/+ 1s. Various factors that may influence delay duration are investigated. Correlations are found between delay length and the duration of the speech sequences associated with the head movements. Effects due to the different head movement types are also discussed.
Index Terms: head movements, movement-speech alignment, delays, dialogues.

1. Background

Many studies have claimed that speech and gesture, in particular hand gestures, are two manifestations of the same underlying cognitive mechanism [2], [3], [4], [5], [6]. One aspect of this tight relation is the temporal coordination between the two modalities. It is generally agreed that hand gestures are coordinated with prosodic events, such as pitch accents and prosodic phrase boundaries [7], [8], [9], [10]. It has also been shown experimentally that subjects are sensible to asynchrony, especially when gesture strokes are made to lag behind the accompanying speech [11], and also that coordination with prosody contributes to the well-formedness of multimodal signals [12].

These studies deal with hand gestures, especially beats. Head movements often have the same quality of manual beats, by being rapid, simple and often repeated movements. Therefore, we would expect them also to show tight temporal synchronisation with the words they co-occur with. Coordination between head movements and speech is discussed in [13], where it is claimed that speakers' head movements are attuned to prosody in establishing peaks and prosodic boundaries especially in cases of high intensity. Furthermore, in [14], it is argued that coordination with speech, together with physical properties of head movements (cyclicity, amplitude, duration) are indicative of the diverse communicative functions of the movements themselves. However, the temporal synchronisation between the two modalities is not described in detail, and the datasets explored in these papers only consist of a couple of hundreds of head movements.

In this paper, we look at temporal synchronisation at the level of onsets and offsets of movements and associated speech, and we analyse a larger dataset.

2. The corpus

The data used in this study come from the Danish NOMCO corpus of first encounter dialogues, a collection of twelve video-recorded dialogues between Danish speakers for a total of about an hour of interaction. The annotation consists of the speech transcription as well as a rather fine-grained annotation of the speakers' gestural behaviour, including their head movements. In addition, each movement is explicitly linked to the speech segment which is semantically associated with it.

Figure 1: Annotation of a head movement in the Danish NOMCO corpus.

For instance, Figure 1 shows the ANVIL [15] annotation board concerning a head movement of type *jerk* (up-nod), which has been linked to the word *okay* in the speaker's own speech stream through the feature *MMRealationSelf*. More detail about the corpus, which is one of a collection of Nordic first encounter dialogues, can be found in [1], and [16].

3. Temporal coordination between head movements and speech

The total number of head movements in the NOMCO corpus is 3117. We are only interested in head movements that are linked to word sequences in the gesturer's own speech stream, and ignore unimodal head movements performed while the interlocutor is speaking. That leaves a subset of 2795 movements, which will be used to analyse movement-speech synchronisation in this study. The duration of most head movements in this dataset is around 1s, although there are occurrences of up to 7s (mean = 0.93s, sd = 0.58s). The duration of the word sequences linked with the head movements, on? the other hand, is on average shorter but with single outliers of 8s and 12s (mean = 0.59s, sd = 0.67s). The distributions are shown in Figure 2.

In what follows we analyse synchrony between head movements and associated speech sequences by looking at start and end delays between the two. A positive start delay means that

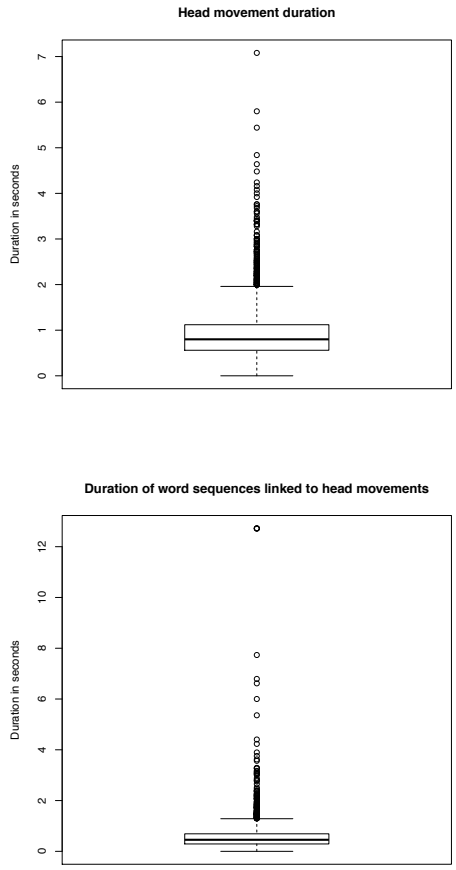


Figure 2: Distributions of the durations of head movements (above) and associated speech sequences (below).

speech onset follows movement onset, in other words that the head movement starts before the associated speech. A positive end delay, on the other hand, means that speech offset follows movement offset, in other words that the head movement ends before speech does.

On average, in our data head movements tend to start 0.05s before the onset of the associated speech sequence (sd = 0.40s), and to end 0.28s after its offset (sd = 0.64s). The histograms in Figure 3 show that in more than 2500 cases (out of the total 2795), delays range between -0.5 and 0.5, and that about 1750 delays are actually positive delays in the range 0 to 1, meaning that in almost two thirds of the cases head movements start before the corresponding speech. Looking at the end delays, on the other hand, we see that slightly more than 1800 are distributed in the range -1 to 0, meaning that in almost two thirds of the cases, the head movement ends up to 1s after speech offset. To have an intuition of what a one second delay means, we can compare it with the mean word duration in the whole NOMCO corpus, which is 0.21s, or the mean length of a linked speech sequence in the dataset, which is as we saw 0.59s. It can also be mentioned that in the already cited study in [11] it is found that subjects are sensible to asynchrony of as little as 0.2 seconds if a gesture lags behind speech, whereas in[12] it is claimed that subjects react to gesture-speech misalignments of at least 0.5 seconds. Thus, a delay of 1s is not negligible, in that it corresponds to four words, or two speech sequences.

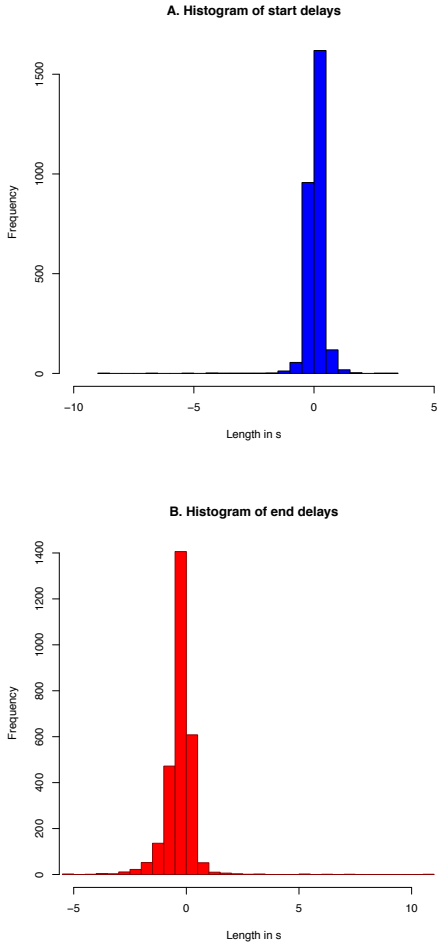


Figure 3: Start and end delays in the NOMCO corpus. In histogram A, bars to the left of zero (negative) correspond to speech preceding the onset of the corresponding movements, and those to the right to speech onset following movement onset. In histogram B, bars to the left of zero count speech ending before, and those to the right speech ending after movement offset. Histogram bins correspond to intervals of half a second.

In the remainder of the paper we will discuss a number of factors which may have an influence on the polarity and duration of the delays.

3.1. Delays in the individual conversations

Some variation can be observed in the individual conversations. In the top graph of Figure 4, which shows means and confidence intervals for start delay duration in the various files, we see that the mean delay duration varies from 0.13 (file M2_M4) to -0.06 (file M6_F1). All the means relating to end delay duration in

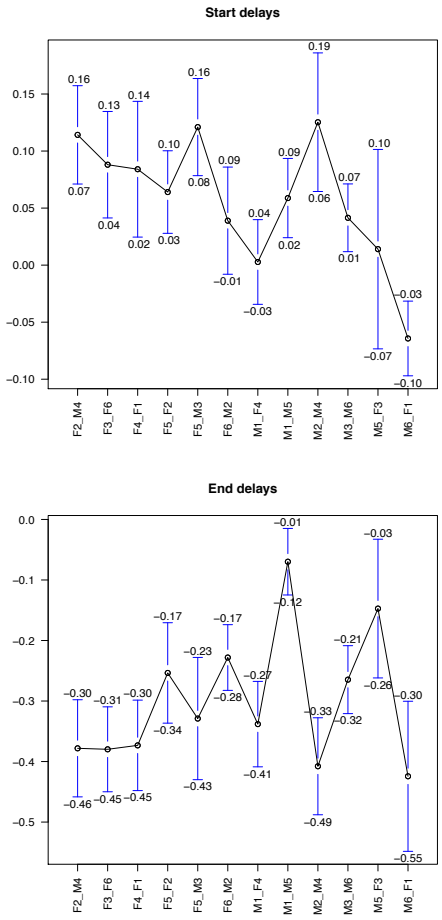


Figure 4: Duration of start and end delays in the twelve NOMCO conversations (means and confidence intervals).

the second plot of the same figure, on the other hand, are in the negative part of the chart (movement ending after speech) and vary from -0.07 (file M1_M5) to -0.42 (file M6_F1).

There is a very small but significant effect of conversation on start delay length (ANOVA, $F(11,2783) = 3.958$, $p<0.001$) and end delay length (ANOVA, $F(11,2783) = 7.387$, $p<0.001$). Only the differences between start delay duration in dialogue M6_F1 and seven of the other dialogues (F2_M4, F3_F6, F4_F1, F5_F2, F5_M3, M1_M5, and M2_M4) reach statistical significance (Tukey's HSD: p-values between < 0.05 and < 0.001). Looking at end delays, on the other hand, statistically significant differences are found between 14 of the pairwise comparisons, all of which involve either dialogue M1_M5 or M5_F3 (Tukey's HSD: p-values between < 0.05 and < 0.001).

3.2. Delays and individual speakers

We also looked at whether speakers differed from each other in their delay durations (Figure 5). We found a very small but significant effect of individual speaker on start delay length (ANOVA, $F(11,2783) = 2.633$, $p<0.001$) and end delay length (ANOVA, $F(11,2783) = 8.708$, $p<0.001$). As far as start delay duration is concerned, only the differences between speakers M4 and M1 on the one hand, and M4 and M6 on the other, reached significance, while 17 of the pairwise comparisons did when looking at end delays (Tukey's HSD).

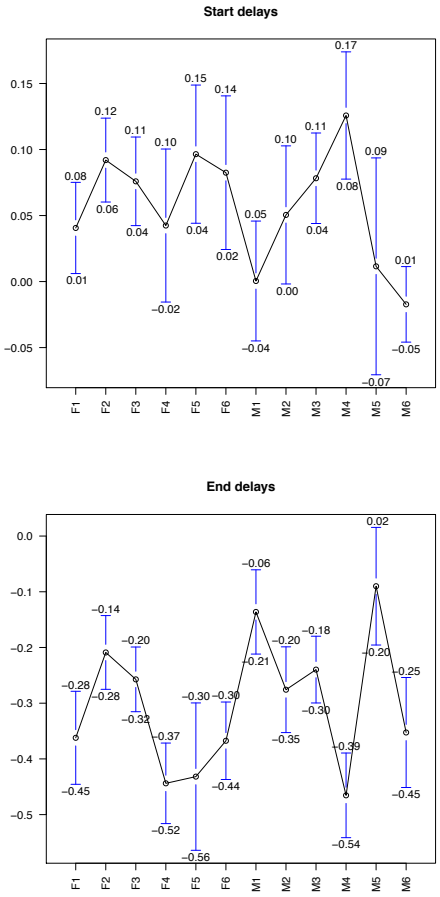


Figure 5: Duration of start and end delays produced the twelve NOMCO speakers (means and confidence intervals).

3.3. Delays and movement type

Start delay duration varies also depending on the type of movement (Figure 6), ranging on average between exact onset synchrony in the case of jerks (upnods) to an average positive delay of 0.12s in the case of waggles. None of the differences, however, reaches statistical significance. If we look at end delays, and leave out the category "Head Other", which collects cases of unclear movements, average duration varies between -0.13 for jerks to -0.54 for waggles. Head movement has in fact a small but significant effect on end delay duration (ANOVA, $F(8,2785) = 11.36$, $p<0.001$). The significant pairwise differences are shown in Table 1.

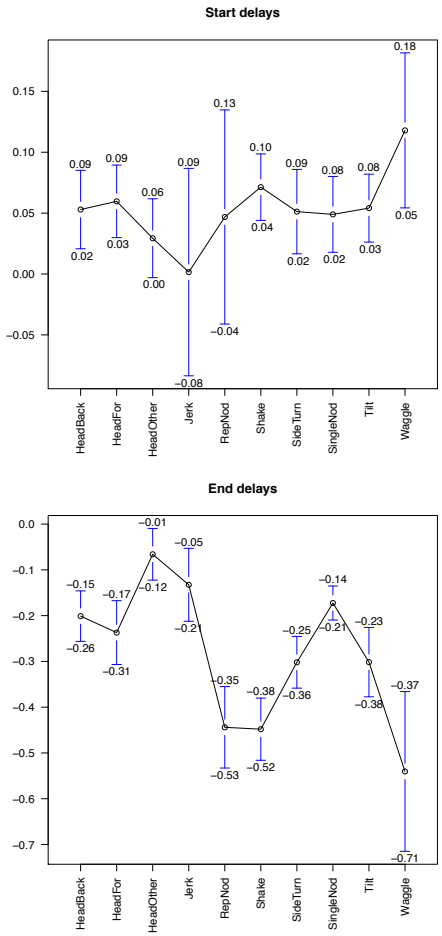


Figure 6: Start and end delays plotted against different head movement type (means and confidence intervals.)

3.4. Effect of movement and speech sequence duration on delays

Finally, the relations between delay length and head movement duration on the one hand, and between delay length and duration of associated speech segments on the other, were also investigated. In both cases, log values were used to diminish the effect of outliers on the correlation coefficient. No correlation was found between delay length and the duration of the head movements. On the other hand, a moderate *negative* correlation can be observed between start delay length and the duration of the speech segments (Pearson's r = -0.57), while a moderate *positive* correlation can be seen between end delay length and speech segment duration (Pearson's r = 0.40). The corresponding plots can be seen in Figure 7. In general, this means that the longer the speech chunk associated with the head movement is, the later the head movement starts and the earlier it ends. Interestingly, the strength of both correlations varies depending on head movement type, as shown in Tables 2 and 3.

We see that there are movement types for which there is a strong correlation between delay and linked speech duration both at onset and offset (jerks, repeated nods) and movement

Table 1: *Significant differences between end delay mean values for different head movement types (Tukey's HSD)*

Pairwise comparison	p value
RepeatedNod-HeadBackward	<0.001
RepeatedNod-HeadForward	<0.01
RepeatedNod-HeadOther	<0.001
RepeatedNod-Jerk	<0.001
Shake-HeadBackward	<0.001
Shake-HeadForward	<0.001
Shake-HeadOther	<0.001
Shake-Jerk	<0.001
SideTurn-HeadOther	<0.001
SingleNod-RepeatedNod	<0.001
SingleNod-Shake	<0.001
Tilt-HeadOther	<0.001
Waggle-HeadBackward	<0.01
Waggle-HeadForward	<0.01
Waggle-HeadOther	<0.001
Waggle-Jerk	<0.001
Waggle-SingleNod	<0.001
Waggle-Tilt	<0.05

Table 2: Pearson's r values showing correlation strength between start delay length and speech chunk duration related to different head movement types.

Head Movement Type	No. of cases	Pearson's r
Jerks	167	-0.94
Repeated nods	327	-0.73
Single nods	244	-0.52
Side turns	417	-0.50
Head other	199	-0.45
HeadB	237	-0.40
Tilts	455	-0.34
Shakes	325	-0.27
HeadF	338	-0.25
Waggles	86	-0.18
All	2795	-0.57

Table 3: Pearson's r values showing correlation strength between end delay length and speech chunk duration related to different head movement types.

Head Movement Type	No. of cases	Pearson's r
Jerks	167	0.81
Head other	199	0.65
Tilts	455	0.60
Repeated nods	327	0.57
Single nods	244	0.42
HeadF	338	0.36
HeadB	237	0.33
Side turns	417	0.25
Shakes	325	0.16
Waggles	86	0.03
All	2795	0.40

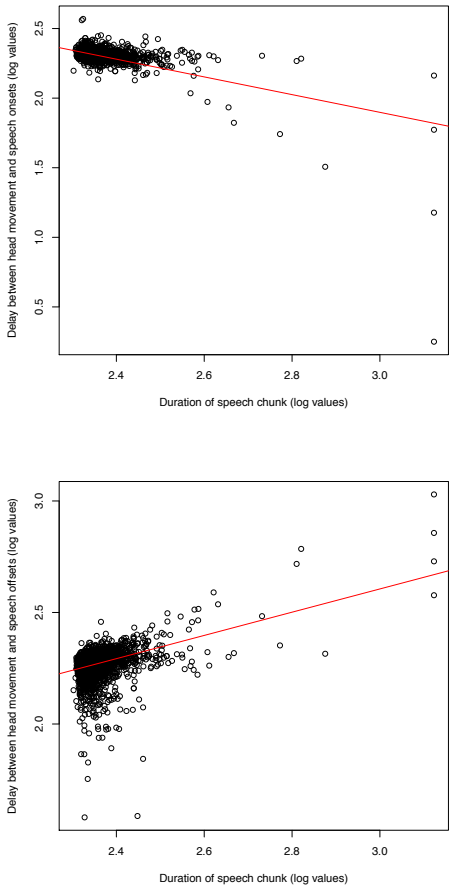


Figure 7: *Correlations between start delays and speech sequence duration (above), and between end delays and speech sequence duration (below). Log values are used.*

types for which the correlation is weak or non-existing at both ends (shakes, waggles). There are, however, also movement types that display a strong correlation between delay and linked speech duration only at onset (side turns, single nods), or only at offset (tilts)[1].

It is tempting to try to make sense of these differences in terms of the relative duration of the various movements, shown in Figure 8. Thus, jerks differ substantially from shakes and waggles both in terms of average duration and duration variance, for example, and they also behave differently in the correlations. Shakes resemble waggles, and likewise they behave very similarly as far as the correlations are concerned. On the other hand, side turns and tilts appear quite similar as far as duration is concerned, and yet they behave in different ways for what concern the correlation between delay and speech duration. In other words, although movement duration may have an effect on the way some head movement types and speech are

[1]The 'Head other' category also shows a strong correlation between end delay and speech offset. However, it is not clear what movement types are grouped in this class.

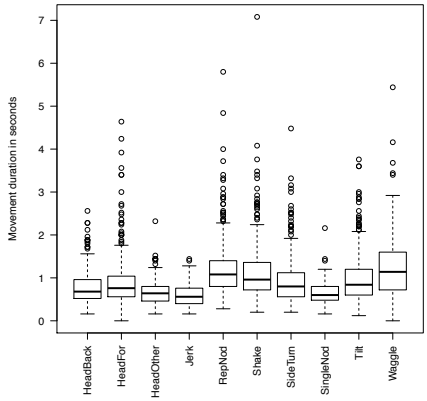


Figure 8: Head movement duration plotted against head movement type.

temporally coordinated, other factors are certainly at play. Examples might be the different kinetic properties of the various movement types, and the alignment between movement strokes and prosodic accents.

4. Conclusion

In a general sense it can be claimed that head movements are temporally synchronised with the associated speech sequences both at movement onset and offset. However, there are delays in both direction in the range of -/+ 1s, which is not a negligible time lag if we consider that subjects have been shown to be sensitive to delays of 0.5s.

Small effects on such variance may be explained in terms of conversation or speaker specific differences. Movement type also has a small effect on the offset delays, where we saw that especially shakes and waggles are responsible for significant differences with respect to other movement types. But the clearest feature that was found in this study was the correlation between delay length and duration of linked speech sequences, which is negative in the case of onset delays, and positive in the case of offset delays. In general, the longer the speech sequence is, the later the movement starts, and the earlier it finishes. This, in turn, can be interpreted as a general tendency for the overlap between head movement and speech sequence to be maximised.

This general pattern, however, varies depending on the movement type, with some types showing a more systematic adherence to the general tendency than others. While these differences seem to be related to the internal duration of the head movement in some cases (jerks, shakes, waggles), duration alone cannot explain the different behaviours of other movement types (e.g. nods, tilts and side turns). A more precise characterisation of the synchronisations patterns for these movement types probably needs to take into account the alignment between movement stroke and prosodic peak, or kinetic features such as amplitude and intensity.

We believe these results are interesting not only in their own right, but also in the context of development of speech production models involving different motional modalities.

5. References

[1] P. Paggio, J. Allwood, E. Ahlsén, K. Jokinen, and C. Navarretta, "The NOMCO multimodal nordic resource - goals and characteristics," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010.

[2] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago, 1992.

[3] ——, *Gesture and thought*. University of Chicago Press, 2005.

[4] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[5] S. Kita and A. Özyürek, "What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking," *Journal of Memory and Language*, vol. 48, no. 1, pp. 16–32, 2003.

[6] J.-P. De Ruiter, "The production of gesture and speech," in *Language and Gesture*. Cambridge University Press, 2000.

[7] D. Bolinger, *Intonation and its parts: Melody in spoken English*. CA: Stanford: Stanford, 1986.

[8] A. Kendon, "Gesture and speech: two aspects of the process of utterance," in *Nonverbal Communication and Language*, M. R. Key, Ed. Mouton, 1980, pp. 207–227.

[9] D. P. Loehr, "Gesture and intonation," Ph.D. dissertation, Georgetown University, 2004.

[10] ——, "Aspects of rhythm in gesture and speech," *Gesture*, vol. 7, no. 2, 2007.

[11] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes*, vol. 26, no. 10, pp. 1457–1471, 2010.

[12] G. Giorgolo and F. A. Verstraten, "Perception of 'speech-and-gesture' integration," in *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, 2008, pp. 31–36.

[13] U. Hadar, T. Steiner, E. C. Grant, and F. C. Rose, "Head movement correlates of juncture and stress at sentence level," *Language and Speech*, vol. 26, no. 2, pp. 117–129, 1983.

[14] U. Hadar, T. Steiner, and F. C. Rose, "The timing of shifts of head postures during conversation," *Human Movement Science*, vol. 3, no. 3, pp. 237–245, 1984.

[15] M. Kipp, *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com, 2004.

[16] C. Navarretta and P. Paggio, "Verbal and Non-Verbal Feedback in Different Types of Interactions," in *Proceedings of LREC 2012*, Istanbul Turkey, May 2012, pp. 2338–2342.

Getting it right:

Advanced Danish learners of Italian acquire speech and gesture L2 forms

Bjørn Wessel-Tolvig

Centre for Language Technology, University of Copenhagen
bwt@hum.ku.dk

Abstract

This paper investigates whether advanced Danish learners of Italian are able to acquire speech and gesture patterns typical of a typologically different target language and consequently move away from patterns typical of their L1. Results show that the Danish learners are able to acquire and use typical Italian lexicalization patterns, but more importantly their L2 speech-gesture co-expressivity reveals that they have reorganized semantic representations and shifted attention towards new types of information.
Index Terms: Second Language Acquisition, motion events, thinking for speaking, gesture, conceptualization.

1. Introduction

Findings within the *Thinking-for-speaking* (TFS) - *Second Language Acquisition* (SLA) paradigm often show very different results in terms of whether learners of a second language (L2) are able to acquire L2 form-meaning pairings and lexicalization patterns, and subsequently shift attention towards new types of information typical of the target language (for recent overviews see:[1, 2]). Nearly all studies highlight the difficulties learners face on the road to acquiring target-like ways of expressing themselves in another language. Learners often continue to be influenced by lexico-semantic and morpho-syntactic structures of their first language (L1) when speaking an L2, a feature known as cross-linguistic influence or transfer. Only a handful of studies find no major evidence of L1 transfer, which in turn is interpreted as a shift in TFS towards new L2 TFS-patterns. But these studies focus only on lexicalization patterns in L1 and L2 speech and say very little about the cognitive functions or conceptualizations at play in native speakers and language learners. We argue, in line with [3], for a methodological shift towards co-verbal behavior (e.g. co-speech gestures) in studies of SLA to 1) better understand linguistic conceptualization of speakers learning another language, and 2) investigate whether acquiring target-like lexicalization patterns also involves a change in *thinking-for-speaking*.

2. Background

We depart from the conceptual domain of motion. Speaking about motion is central to human communication and all languages have lexical means for describing it. But speakers of different languages show striking variation of how semantic features of a motion event, e.g. path (directionality) and manner (the way movement is carried out) are mapped onto linguistic surface forms. This has lead Talmy [4, 5] to propose a two-way typological classification of languages depending on how the main constituent, PATH OF MOTION, is framed across languages. Speakers of *verb-framed languages* (Romance, Japanese, Semitic) often express path in the main verb (e.g. *ascend*, *enter*) as in (1), whereas speakers of *satellite-framed languages* (e.g. Indo-European - except Romance – Slavic) mainly express path outside the main verb in adverbials or PPs (e.g. *up*, *down*, *into*, *out of*) as in (2). As seen in the examples, the allocation of path also has consequences for expressing manner of motion. Since the main verb is occupied by the expression of path in verb-framed languages, manner must be subordinated in PPs, gerunds or subordinate clauses, if not omitted altogether. In satellite-framed languages the main verb slot is left open to express manner.

(1) La botella **entrò** a la cueva (*flotando*)[1]
'The bottle enters the cave (*floating*)'
(2) The bottle floats **into** the cave

Based on the differences in lexicalization of manner and path, Berman & Slobin [6] have examined whether and to what extent the typological variation among different languages has an effect on speakers' conceptualization, and expression, of motion events.

2.1. Thinking-for-speaking about motion

Slobin hypothesizes that in the process of speaking, experience is filtered through language into verbalized events, what he calls *thinking-for-speaking* [7, 8]. Studies on language diversity and TFS explore how speakers of different languages select and organize information, e.g. about path and manner, depending on the morpho-syntactic possibilities (and constraints) provided by their particular language. The TFS hypothesis thus centers on the effect of language on the cognitive processes during speaking. Native speakers are from childhood (L1 acquisition) trained to pay attention to specific aspects of a motion event, which leads to language-specific rhetorical styles in the way speakers not only speak about path and manner, but also the amount of attention paid towards them. Therefore, if differences in lexicalization across languages give rise to cross-linguistic differences in cognition, it can have important implications for SLA [9].

2.2. Thinking-for-speaking in L2

Learning another language not only entails learning new form-meaning pairings, it also involves selecting and

[1] Standard textbook examples by Talmy [4, pp. 69]

(syntactically) organizing information in target-like ways. The process of SLA therefore involves restructuring existing conceptual categories [10] and learning new ways of TFS [11]. This can be a challenging task for L2 learners as L1 patterns learned in childhood seem "*resistant to restructuring in adult second language acquisition*" [8, pp. 89]. Learners may learn new L2 forms, but apply them from an L1 perspective [12].

One of the major questions in SLA is whether L2 learners can overcome the constraints imposed by their native language and learn to conceptualize motion in ways typical of the target language. Studies show that even advanced language learners have problems in reorganizing semantic components and shift attention towards new ways of thinking (for speaking) [10, 13]. Only a handful of studies find little evidence of L1 transfer and suggest restructuring of semantic representations are possible. In a series of studies of Danish (satellite-framed language) L2 learners of Spanish (verb-framed language), Cadierno and colleagues find that learners have no problems acquiring typical Spanish constructions, but still show traces of L1 transfer by more elaborate descriptions of path using redundant path particles (adverbials) with path verbs, a term coined '*satellization*' by Cadierno [14]. However, no transfer of dominant Danish manner verb + PP structure is visible.

Bernini et al. [15] however argue that L1 background alone cannot clearly predict L2 behavior. Comparing German and English learners (both satellite-framed languages) of Italian (verb-framed), the learner groups showed different lexicalization patterns in L2 Italian. English L2 speakers acquired Italian verb-framed patterns, whereas the German speakers did not fully master this. They used more light path verbs (deictic) like *andare* – 'to go', fewer path only verb constructions and preferred to lexicalize path in PPs. Although studies often find L1 transfer, or traces of general language learner behavior, some argue that learners can develop appropriate L2 TFS patterns over time [16].

One factor all of the abovementioned studies have in common is that they focus on speech alone. In a critique, Athanasopoulos & Bylund point out that many of these studies investigating TFS "*do not in fact provide sufficient data on thought processes during speaking, but only describe linguistic diversity in the sense that they report typologically-constrained verbalizations produced by speakers of different languages*" [3, pp. 95]. That is, the studies may provide detailed linguistic analyses on information structure in L1 and L2, which is an essential starting point for investigating language diversity, but they reveal little about online linguistic conceptualization. Athanasopoulos & Bylund, among others, argue that looking at co-verbal data (ERP, eye-tracking, nonverbal tasks, gesture) might reveal more about cognitive processes during speaking than speech alone.

2.3. Why gestures?

We often gesture when we speak. Gestures are semantically and temporally tightly related to speech and language, and are seen as "*forming an 'integrated' system which is planned and processed together*" [17, pp. 78]. Because of the tight semantic and temporal relationship, co-speech gestures are influenced by information structure, that is: what type of information is selected for speech, and how the information is linguistically organized. Several cross-linguistic studies show that speakers of different languages not only encode and express meaning differently, but also distribute gestures differently when re-narrating the same storylines [18]. Speakers of verb-framed languages, who often need two verbal clauses to express manner and path, also tend to divide manner and path into two separate gestures: one for manner, one for path. Speakers of satellite-framed languages, on the contrary, often express manner and path within one clause and consequently produce one gesture conveying information about path alone or conflate manner and path into one single gesture [19]. The co-expressivity of meaning in speech and gesture indicates that they are conceptually linked and may as such reflect how events are conceptualized. The speech-gesture co-expressivity is therefore interesting for studies of TFS in SLA. Co-speech gesture may be used to investigate whether, and to what extent, L2 learners are able to reorganize semantic representations from their native language onto an L2 with different types of representations.

If learners acquire target-like representations of motion, their co-speech gestures should reflect this in target-like gesture patterns. Studies investigating L2 speakers' speech and gesture patterns, and acquisition of such, mainly find 1) difficulties for learners in expressing motion in target-like ways both in speech and gesture, and thereby seem to retain L1 TFS patterns [20, 21], 2) properties of both source and target language in L2 production [22], and thus that a shift towards L2 TFS *is* possible for some aspects of motion [23, 24], 3) evidence of restructuring of representations exemplified in typical speech and gesture L2 forms [16, 25]. Stam [24] finds that L2 learners' gestures reveal L1-based TFS with fluent L2 speech, but subsequently show that over a period of 14 years exposure to the target-language, *one speaker's* speech and gesture patterns shift towards L2 typical patterns. Özyürek [25] find that very advanced Turkish learners of English, being submerged into target-language culture for more than 10 years, acquire typical L2 speech forms which is also reflected in L2-like gesture patterns.

We might assume that advanced learners, who are grammatically correct and fluent in the L2, have restructured semantic representations and acquired L2 TFS-patterns. Evidence from co-speech gestures, however, questions such assumptions. If L2 speakers' gestures show L1-based gesture with L2-fluent speech instead of L2-typical gesture form and distribution, then we can hypothesize that learners have not fully re-conceptualized motion events and shifted attention towards new information structures.

2.4. Present study: Motion in Danish and Italian

Danish and Italian represent two typologically different languages. Danish is categorized as a prototypical satellite-framed language as path of motion is almost exclusively expressed in verb particles (adverbs, PPs) and manner is likewise most often expressed in the main verb [26, 27] as in (3) where the path particle is in boldface.

(3) Bolden ruller **ind** i huset
'Ball.the rolls in-to house.the'

Italian is, along with the other Romance languages, considered a verb-framed language, as path is often expressed in the main verb and manner subordinated in adverbial manner expressions or subordinate clauses as in (4).

(4) Il pallone **entra** nella casa (*rotolando*)
'The ball enters in.the house (*rolling*)'

But lexicalization patterns are not fixed. Most languages possess a variety of different means of expressing motion [28]. In fact, verb-particle constructions are allowed in Romance languages (at least in non-boundary-crossing situations), also in combination with manner verbs as in (5).

(5) Il Pallone rotola **giù** per la strada
'The ball rolls down on the street'

In light of recent research into verb-particle constructions, Italian may be different in respect to other verb-framed languages in the colloquial and frequent use of verb-particle constructions [29]. This has led Talmy to (*re-*)classifying Italian as a language with 'split system' possibilities [5]. Although verb-particle constructions are frequently used in Italian, the preferred way of lexicalizing motion is based on a verb-framed schematic with path in the main verbal clause +/- subordinated manner expressions. Recent studies of Italian speech-gesture patterns show that when Italian speakers divide path and manner in speech, they typically also produce two separate gestures. However, when they conflate manner and path in a verb-particle construction, they produce one gesture [30, 31].

2.5. Research question

The question is whether and how deep semantic preferences from the L1 may influence conceptual representations in L2, and whether such representations can be restructured towards target-language representations. We look at co-speech gestures to investigate L1-based thinking in otherwise fluent L2 production or indications of a change in TFS towards the L2.

3. Method

3.1. Participants

A total of 10 speakers participated in the study: 5 native Italian speakers (female 4, Mean age 29.4, SD 7.05), all grad students of Roma Tre (Rome, Italy) and 5 highly advanced Danish learners of Italian (female: 4, mean age 36, SD 6.71), all post-grads, Masters of Arts in Italian Language and Literature from the University of Copenhagen and Copenhagen Business School. In a language proficiency test (cloze test) they scored a mean 91.38% correct (SD 3.28) and in a self-rated language background questionnaire on proficiency they scored a mean 4 (SD 0.72) of 5 – 5 is best). All had lengthy experience living in the target language culture (mean 19.6 months/SD 12.52) and to some extent spoke Italian on a weekly basis (3.8 hours/SD 4.82). The Danish speakers represented both the Danish L1 and L2 Italian group.

3.2. Experimental design

The participants individually watched 16 short cartoon videos (8 fillers) consisting of material from The Tomato Man Project [32] and from Boundary Ball [33] involving animated figures, a tomato or triangle, jumping or rolling up and down a hill or into and out of a house. The participants narrated the events to a confederate native listener with the instruction that a third naïve listener would watch the recordings and be able to understand and elaborate on the details of the storylines based on their narrations. The order in which the Danish L1 and L2 speakers were tested was counterbalanced.

3.3. Encoding

Speech was tokenized and the events categorized as to how manner and path was expressed syntactically in the narration of target events. Narrations with manner verbs + a path denoting *satellite* (e.g. adverbs, preposition) were categorized as 'one-clause' manner-path conflating constructions (MP). Constructions containing only manner or only path were labelled MO and PO respectively, and constructions containing both manner and path in two separate (verbal) elements (e.g. path verbs + subordinate manner expressions or subordinate clauses) were categorized as a 'two-clause' construction (PO+MO). Examples and labels can be seen in Table 1. Although the Italian adverbial gerund form may not *per se* be categorized as a subordinate clause in itself, we categorize expressions with a path verb and a gerundive manner expression as a '*two clause*' construction type, since gerunds in previous studies often are dealt with as subordinated manner elements outside the main verbal phrase, and because they may constitute a processing unit in themselves similar to subordinated manner expressions in subordinated clauses.

Clause type	Example	Labels
One clause	And he **rolls up** the hill	MP
One clause	He **jumps into** the house	MP
One clause	The tomato **rolls**	MO
One clause	He **descends** the hill	PO
One clause	It **goes down** the hill	PO
Two clauses	He **descends** while **rolling**	PO+MO
Two clauses	He **enters** the house **jumping**	PO+MO

Table 1: Speech constructions, examples and labels

Gestures were categorized as to what information the co-speech gesture contained as seen in Table 2. The label 2G is given to gesture constructions in which two separate gestures are expressed within a target event, e.g. one for path and one for manner.

Gesture type	Representation	Labels
Path	Representing only the path of motion with no explicit reference to manner	PG
Manner	Depicting only the manner of motion, that is how the figure moves, with no indication of the path	MG
Manner-path conflating	Conflating both the manner and the path of motion into one single gesture	MPG
Two separate gestures	Two separate gesture containing manner and path information	2G

Table 2: Gesture examples and labels

4. Analysis

The participants produced a total of 137 motion events and a total of 180 gestures. In what follows we present a quantitative analysis of speech patterns between the three groups of speakers and an analysis of speech-gesture co-expressivity. We use within-group repeated measures ANOVA for the Danish-L2 group (as both groups contained the same speakers), and a between-group factorial mixed effects ANOVA for the Danish-Italian and Italian-L2 group. We subsequently performed Bonferroni post-hoc tests. For path only (PO) and separate clause constructions (PO+MO) and

gesture type we only carried out statistical analysis between the Italian and L2 speakers since the Danish speakers did not produce any of these clause constructions.

4.1. Clause type constructions

Figure 1 visualizes the lexicalization patterns most frequently used by the three groups. The plot shows Danish L1 speakers' exclusive use of a tight manner verb + path satellite configuration (MP), which is in line with previous studies of Danish lexicalization patterns [14]. Furthermore, the pattern confirms Danish to be a rigid and prototypical satellite-framed language with few possibilities for lexicalizing motion in different ways.

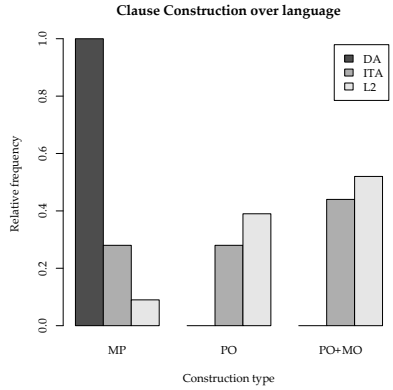


Figure 1: Relative frequency of Clause Construction in the three Language groups

The data for Italian confirms its rather particular position as a typological 'split' system language. In 28% of narrations manner is expressed in the main verb and path in a satellite to the verb (MP). However, a total of 72% of narrations are construed by means of typical verb-framed patterns with path expressed in the main verbal clause +/- subordinate manner expressions. Although Italian speakers have at their disposition conceptually lighter manner verb + particle constructions, the speakers overall frame motion in standard verb-framed fashion, especially in 'two clause' constructions with subordinated manner expressions (44%).

Generally, Italian speakers in this study do not omit manner; rather it is expressed extensively both through main verb + satellite and subordinated constructions. Given that Italian speakers can frame manner and path within one clause, why is this pattern not more widespread? One explanation is the boundary-crossing constraint which hinders speakers of verb-framed languages from expressing path in a satellite (in conjunction with a manner verb) when a figure crosses a spatial boundary [34]. Another explanation is that lexicalization patterns are so deeply entrenched in human cognition that the Italian speakers may prefer heavier constructions (PO+MO) to lighter constructions (MP or PO) to fulfill verb-framed schematics also including manner.

The Danish L2 Italian speakers, whose L1 lexicalization patterns clearly are grounded in a preference for tight manner verb + path satellite construction, must reorganize semantic representations and move away from mapping manner in the main verb and path in a satellite to fit standard target-like patterns. The L2 learners succeed in expressing motion in ways similar to the target language, and well distanced from their L1 patterns. They succeed in suppressing manner in the main verb (9%) and produce typical target-like constructions with path only (PO) and path only constructions + manner subordination (PO+MO) (91%). When comparing the Danish-Italian group there is a main effect of Language ($F(1,8) = 6930 < 0.001$) and a main effect of Construction ($F(2,8) = 65 < 0.001$) and an interaction between the two ($F(2,8) = 78 < 0.001$). Between the Danish-L2 group there is a main effect of Language ($F(1,4) = 1047 < 0.001$), a main effect of Construction ($F(2,4) = 28 < 0.001$) and an interaction between the two ($F(2,4) = 78 < 0.001$). Bonferroni post-tests showed significant pairwise differences between construction types and languages. Between the Italian-L2 group we find no effect of Language ($F(1,8) = 3.71 = 0.09$), only a marginal effect of Construction ($F(2,8) = 3.64 = 0.049$) and no interaction between the two ($F(2,8) = 1.28 = 0.3$). A Bonferroni post-test shows a marginal pairwise difference in expressing MP (0.03), but no difference in PO and PO+MO between the groups.

The statistical analysis for clause type constructions overall shows that Danish and Italian speakers express motion in significantly different ways. The L2 speakers express motion in ways significantly different from their L1, and these expressions in L2 are not significantly different from L1 Italian. The results for speech also show that the L2 speakers lexicalize motion in a more standard verb-framed fashion than the L1 Italian speakers producing fewer manner verb + path particle constructions, a pattern present in colloquial Italian. From the speech data we could infer that the L2 learners had restructured semantic representations. However, co-speech gestures may give a clearer view on linguistic conceptualization of the events in L2.

4.2. Clause type constructions and gestures

Looking at speech and gesture combined, we investigate how the packaging of manner and path in speech constructions is reflected in co-speech gesture. We divide the bar plots by language for reasons of simplicity, but statistical analysis is carried out within clause constructions and gesture types. Figure 2, 3 and 4 visualize how gestures are combined with speech constructions within the three language groups.

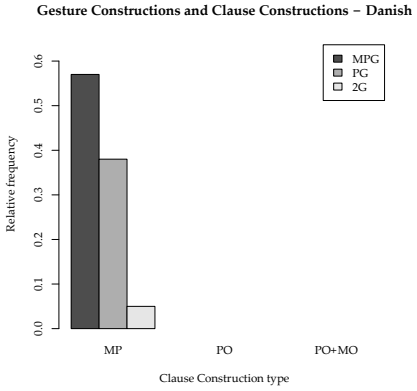


Figure 2: Relative frequency of Gesture Constructions over Clause Construction – Danish

Figure 2 shows that Danish L1 speakers almost exclusively produce manner-path conflated gestures (MPG) and path only gestures (PG) combined with their exclusive use of one clause manner-path constructions (MP). This speech-gesture pattern supports previous research for satellite-framed languages [19] showing that when speakers linguistically conflate manner and path in one single clause (or processing unit following Kita & Özyürek and colleagues), they also produce one single gesture which may either represent the same information as speech (MPG) or downplay manner in gesture (PG) possibly due to high presence in speech.

Figure 3 shows Italian L1 speakers mainly use manner-path conflated gestures (MPG) or path only gestures (PG) when expressing manner and path within a single clause (MP). Moreover, they tend to produce path only gestures with path only constructions (PO) and when separating manner and path in 'two clauses' (PO+MO), also separate manner and path in gesture (2G). The division of manner and path in gesture with the 'two clause' constructions reflect the conceptual separation of manner and path as pertaining to two separate processing units [19]. The division of gestures is also found in other studies for Italian [30] and other verb-framed languages [See:18].

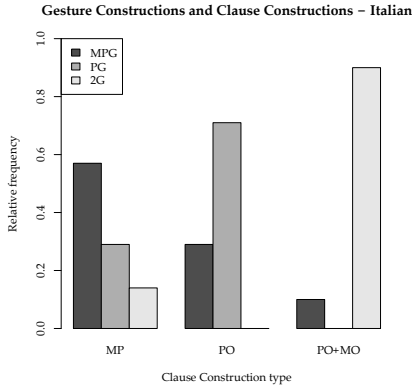


Figure 3: Relative frequency of Gesture Constructions over Clause Construction - Italian

For the L2 learners to fully acquire gesture patterns in L2 Italian, they must resemble the Italian speech-gesture patterns expressing path gestures with path constructions and two separate gestures when separating manner and path in speech.

Figure 4 illustrates how the L2 speakers align path gestures (PG) with path constructions (PO) and produce two separate gestures (2G) when separating manner and path in speech (PO+MO). Although frequencies for (MP) may seem high, remember that L2 speakers only produced manner verb + path satellite constructions in 9% of all narrations which with gesture production amount to 4 observations within the MP category in total.

The findings show that the learners not only acquire target-like speech patterns, but also use target-like gesture when expressing motion. The speech-gesture synchrony may suggest a restructuring of semantic representations from a rigid satellite-framed system towards a more standard verb-framed Italian.

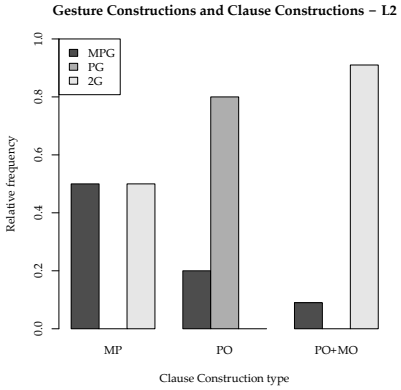


Figure 4: Relative frequency of Gesture Constructions over Clause Construction – L2

4.2.1. Manner-path conflated constructions

When comparing the Danish-Italian speakers within the MP constructions we find a main effect of Language ($F(1,8) = 115 < 0.001$), a main effect of Gesture ($F(3,8) = 12 < 0.001$) and an interaction between the two ($F(3,8) = 4.6 < 0.005$). Between the Danish-L2 speakers there is a main effect of Language ($F(1,4) = 509 < 0.001$), a main effect of Gesture ($F(3,4) = 7.4 < 0.005$) and an interaction between the two ($F(3,4) = 12 < 0.001$). Between the Italian-L2 speakers there is a marginal effect of Language $F(1,8) = 6.59 < 0.05$) no effect of Gesture ($F(3,8) = 2.39 = 0.09$) and no interaction between the two ($F(3.8) = 1.43$, $p = 0.26$). Bonferroni post-testing shows a pairwise difference between the Danish-Italian speakers only for MPG (0.003), for the Danish-L2 speakers only for MPG (< 0.001) and PG (0.01) and for the Italian-L2 speakers only for MPG (0.043).

For speech-gesture relationship within the MP category we observe a variation between the three groups. Even when using the same tight one-clause manner verb + path satellite construction, we see a difference in what type of gesture speakers produce. Although we might assume that producing the same type of construction (MP) would result in the same type of gesture distribution across the languages (MPG or PG), it is not entirely the case. Few gesture observations, especially in the L2 dataset within the MP construction, could bias the statistics.

4.2.2. Path only constructions (L2-ITA)

For the path only (PO) constructions there is a marginal effect of Language ($F(1,8)$ 6.59, $p < 0.05$), no effect of Gesture ($F(3,8) = 2.39$, $p = 0.09$) and no interaction ($F(3,8) = 1.43$, $p = 0.25$). A Bonferroni post-test shows pairwise difference only for MPG (0.04). There is a marginal effect of Language between the L2 and Italian L1 speakers, but generally the learners achieve mapping path gestures onto path only constructions in target-like ways with negligible differences.

4.2.3. Path+Manner separated constructions (L2-ITA)

For the PO+MO constructions there is no effect of Language ($F(1,8) = 0.57$, $p = 0.47$), a main effect of Gesture ($F(3,8) = 17$, $p < 0.001$), but no interaction ($F(3,8) = 0.46$, $p = 0.72$).

A Bonferroni post-test show a pairwise difference only for MPG (0.04). The L2 speakers acquire and use two separate manner and path gestures when separating manner and path in speech. The data shows, at least for the 5 Danish learners of Italian, a shift towards a more uniform verb-framed Italian.

5. Discussion

Learning to map semantic features of motion onto new L2 linguistic forms is notoriously difficult for L2 learners, because the L1 patterns are deeply rooted in conceptualization. This study indicates that a restructuring of semantic representation is possible. The findings support previous studies by Özyürek [25] and Stam [16], who also find that advanced language learners' gestures reflect the learners' acquired target-like speech patterns. The learners in this study are highly advanced learners of Italian, they have learned Italian in formal ways during university studies, they have all been submerged into target language culture, and have been in social and linguistic contact with native Italian speakers. Their acquisition of more standard verb-framed forms exemplified in a higher production of PO+MO (than the L1 Italian speakers) may be attributed to formal textbook schooling through university learning combined with a continued L1 focus on expressing manner.

The learners do not overtly use the MP construction in L2 production although the construction is valid in non-boundary-crossing situations. This could indicate minimal transfer from L1 MP rigid patterns, but also that the learners are aware that expressions of directionality are associated with the verb and not explicitly with the particle (satellite). Another factor could be that even advanced learners stick to known formal structures and do not 'play' much with language variation risking ambiguity and being misunderstood.

Looking at gestures, we see that a conceptual shift towards target-like, and even standard target-language, is possible for very advanced learners [25]. The learners are able to reorganize semantic representations towards an Italian TFS-pattern for the domain of motion. Since gestures reflect linguistic conceptualizations, we see that the L2 speakers succeed in aligning path gestures with path only expressions and separate manner and path in gesture when separating manner and path in speech, two types of speech-gesture patterns not found in their L1.

We are careful not to conclude that these data show a full shift in TFS towards target-language patterns. The thinking-for-speaking hypothesis deals with much more than just allocation of manner and path of motion e.g. definiteness and aspect. Moreover, this study is limited to 5 advanced learners with a limited set of motion constructions and verb variation. Despite this, we continue to argue for a 'multimodal approach' to studying Thinking for Speaking [3] and Second Language Acquisition [35].

6. Acknowledgements

Thanks to Patrizia Paggio, University of Copenhagen for feedback, Isabella Poggi and Laura Vincze, Roma 3 University for help with data collection and Francesca Gregorino, University of Torino for annotation. This research was funded by the Danish Council for Independent Research (DFF).

7. References

[1] Z. Han and T. Cadierno, *Linguistic relativity in SLA: Thinking for speaking*. Bristol: Multilingual Matters, 2010.

[2] A. Pavlenko, *Thinking and speaking in two languages*. Clevedon, UK: Multilingual Matters, 2011.

[3] P. Athanasopoulos and E. Bylund, "The 'thinking' in thinking-for-speaking: Where is it?," *Language, Interaction & Acquisition*, vol. 4, pp. 91-100, 2013.

[4] L. Talmy, "Semantics and syntax of motion," in *Language typology and syntactic description, Vol. 3, Grammatical categories and the lexicon*. vol. 3, T. Shopen, Ed., ed Cambridge: Cambridge UniversityPress, 1985, pp. 57-149.

[5] L. Talmy, *Toward a Cognitive Semantics: Typology and Process in Concept Structuring* vol. II. Cambridge: The MIT Press, 2000.

[6] R. A. Berman and D. I. Slobin, *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Psychology Press, 1994.

[7] D. I. Slobin, "Thinking for speaking," in *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 1987, pp. 435-444.

[8] D. I. Slobin, "From ''thought and language'' to ''thinking for speaking'," in *Rethinking linguistic relativity*, J. J. Gumperz and S. C. Levinson, Eds., ed Cambridge: Cambridge University Press, 1996, pp. 70-96.

[9] E. Bylund and P. Athanasopoulos, "Introduction: Cognition, Motion Events, and SLA," *The Modern Language Journal*, vol. 99, pp. 1-13, 2015.

[10] S. Jarvis and A. Pavlenko, *Crosslinguistic influence in language and cognition*. New York: Routledge, 2008.

[11] T. Cadierno, "Motion in Danish as a Second Language: Does the Learner's L1 Make a difference," in *Linguistic Relativity in SLA: Thinking for Speaking*, Z. Han and T. Cadierno, Eds., ed Bristol: Multilingual Matters, 2010.

[12] E. Kellerman, "Crosslinguistic influence: Transfer to nowhere?," *Annual Review of Applied Linguistics*, vol. 15, pp. 125-150, 1995.

[13] P. Larrañaga, J. Treffers-Daller, F. Tidball, and M.-c. G. Ortega, "L1 transfer in the acquisition of manner and path in Spanish by native speakers of English," *International Journal of Bilingualism*, vol. 16, pp. 117-138, March 1, 2012 2012.

[14] T. Cadierno, "Expressing motion events in a second language: A cognitive typological approach," in *Cognitive linguistics, second language acquisition and foreign language pedagogy*, M. Achard and S. Neimeier, Eds., ed Berlin: Mouton de Gruyter, 2004.

[15] G. Bernini, L. Spreafico, and A. Valentini, "Acquiring motion verbs in a second language: The case of Italian L2," *Linguistica e Filologia*, vol. 23, 2006.

[16] G. Stam, "Changes in Thinking for Speaking: A Longitudinal Case Study," *The Modern Language Journal*, vol. 99, pp. 83-99, 2015.

[17] M. Gullberg, "Methodological reflections on gesture analysis in second language acquisition and bilingualism research," *Second Language Research*, vol. 26, pp. 75-102, January 1, 2010 2010.

[18] M. Gullberg, "Thinking, speaking and gesturing about motion in more than one language," in *Thinking and*

speaking in two languages, A. Pavlenko, Ed., ed Bristol: Multilingual Matters, 2011, pp. 143-169.
[19] S. Kita and A. Özyürek, "What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking," *Journal of Memory and Language,* vol. 48, pp. 16-32, 2003.
[20] E. Kellerman and A. M. van Hoof, "Manual accents," *International Review of Applied Linguistics,* vol. 41, pp. 251-269, 2003.
[21] E. Negueruela, J. P. Lantolf, S. R. Jordan, and J. Gelabert, "The 'Private Function' of Gesture in Second Languages Communicative Activity. A Study on Motion Verbs and Gesturing in English and Spanish," *International Journal of Applied Linguistics,* vol. 14, pp. 115-159, 2004.
[22] A. Brown and M. Gullberg, "Bidirectional crosslinguistic influence in l1-l2 encoding of manner in speech and gesture: A Study of Japanese Speakers of English," *Studies in Second Language Acquisition,* vol. 30, pp. 225-251, 2008.
[23] S. Choi and J. P. Lantolf, "Representation and embodiment of meaning in l2 communication: Motion Events in the Speech and Gesture of Advanced L2 Korean and L2 English Speakers," *Studies in Second Language Acquisition,* vol. 30, pp. 191-224, 2008.
[24] G. Stam, "Thinking for speaking about motion: L1 and L2 speech and gesture," *International Review of Applied Linguistics*, vol. 44, pp. 143-169, 2006.
[25] A. Özyürek, "Speech-gesture relationship across languages and in second language learners: Implications for spatial thinking and speaking," presented at the Proceedings of the 26th annual Boston University Conference on Language Development, Somerville; MA, 2002.
[26] M. Jessen and T. Cadierno, "Variation in the categorization of motion events by Danish, German, Turkish, and L2 Danish speakers," in *Variation and change in the encoding of motion events*, J. Goschler and A. Stefanowitsch, Eds., ed Amsterdam: John Benjamins, 2013.
[27] B. Wessel-Tolvig, "Up, down, in & out: Following the Path of motion in Danish and Italian," presented at the Proceedings of the 1st European Symposium on Multimodal Communication, Valletta, Malta, 2014.
[28] J. Beavers, B. Levin, and S. Wei Tham, "The typology of motion expressions revisited," *Journal of Linguistics,* vol. 46, pp. 331-377, 2010.
[29] D. I. Slobin, "The many ways to search for a frog: Linguistic typology and the expression of motion events," in *Relating events in narrative: Typological and contextual perspectives* S. Strömqvist and L. Verhoeven, Eds., ed Mahwah, NJ: Lawrence Erlbaum Associates, 2004, pp. 219-257.
[30] F. Cavicchio and S. Kita, "Gestures Switch in English/Italian Bilinguals," in *MMSYM2014*, Tartu, Estonia, 2014.
[31] B. Wessel-Tolvig, "Breaking boundaries: How gestures reveal conceptualization of boundary-crossing in Italian," presented at the Proceedings of Gespin 4, Nantes, France, 2015.
[32] A. Özyürek, S. Kita, and S. Allen, "Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures.," ed. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, Language and Cognition group, 2001.
[33] B. Wessel-Tolvig, "Boundary Ball: An animated stimulus designed to elicit motion with boundary crossing situations.," ed. University of Copenhagen, 2013.
[34] J. Aske, "Path predicates in English and Spanish: A closer look," in *Proceedings of the Berkeley Linguistic Society 15*, 1989.
[35] A. Brown, "Universal Development and L1–L2 Convergence in Bilingual Construal of Manner in Speech and Gesture in Mandarin, Japanese, and English," *The Modern Language Journal,* vol. 99, pp. 66-82, 2015.