Speech and gestures

Computational Linguistic Studies

Costanza Navarretta
Centre for Language Technology
Department of Nordic Studies
and Linguistics

CLARIN Annual Conference, Pisa
2018

UNIVERSITY OF COPENHAGEN

CLARIN-DK

Content

- Introduction: researchers involved and some terminology
- An introduction to the research on multimodality at the Centre for Language Technology (NorS-UCPh)
- Gesture types and their interpretation
- Selected studies investigating the relation between speech and gestures

RESEARCHERS AND INTRODUCTION

- Colleagues : Patrizia Paggio, University of Copenhagen and University of Malta, Bart Jongejan, Manex Aguirrezabal, University of Copenhagen
- former PhD fellow Magdalena Lis
- Danish annotators mm.: Sara Andersen, Josephine Bødker Arrild, Anette Studsgård, former PhD fellow Bjørn Nicola Wessel-Tolvig, Philip Diderichsen

- Jens Allwood and Elisabeth Ahlsén, University of Gothenburg
- Kristiina Jokinen, University of Helsinki, now AISIT, Japan

Some terms

- A modality is connected to a sense: speech is connected to hearing, gestures, e.g. head movements, facial expressions, hand gestures, body posture, are connected to sight etc.

- Humans interact with the world multimodally. Cognitive studies have determined how different modalities are prioritised/or combined by humans:

  usually visual > audio

What?

- A roaring cat is still a cat

Auditive and visual modalities in communication

Speech and communicative gestures are interpreted together, and they

- are related temporally and semantically (Kendon 2004, McNeill 2005),
- are two manifestations of the same underlying concept (McNeill 1992, 2004, Cassell, McNeill and McCullough 1999, Kelly et al. 1999, Kendon 2004, Kita 2009)

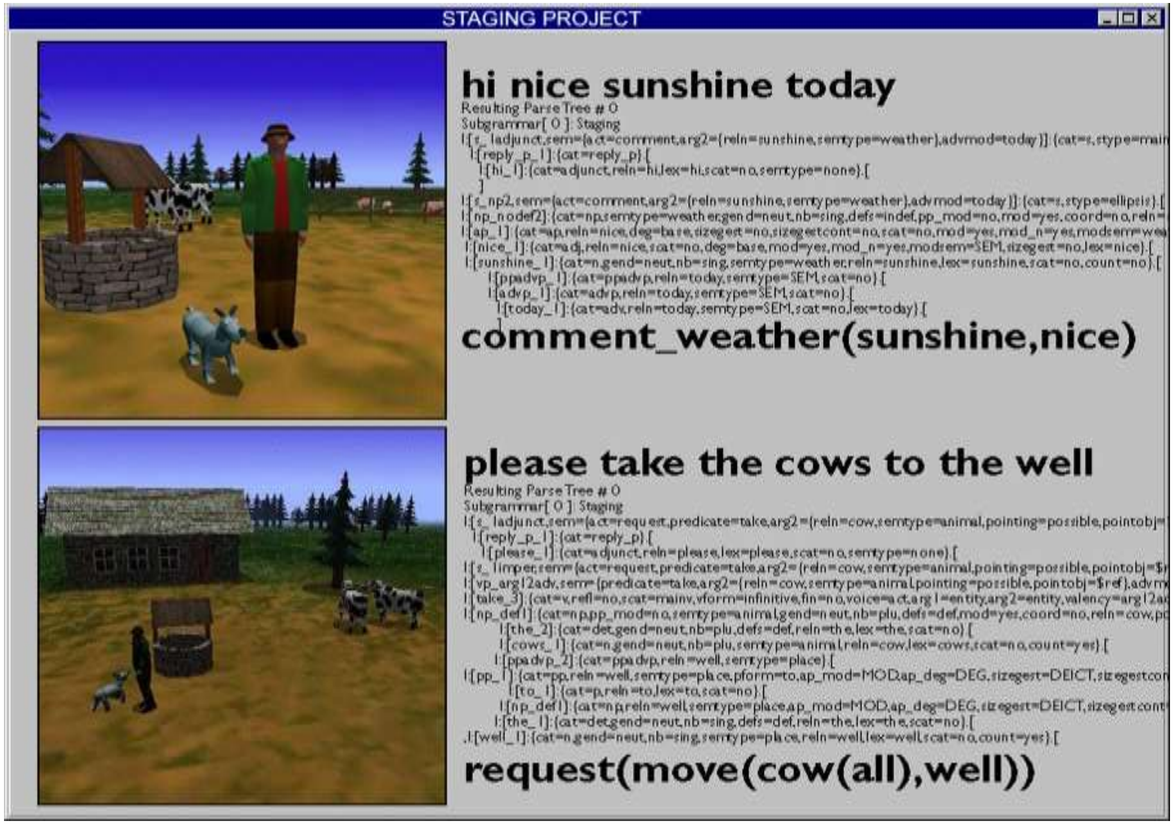
Speech and gestures

- Gestures are multifunctional.
- How we speak and gesture depends on the communicative settings: the communicative situation, the language and culture, the degree of familiarity of the participants, their age, gender, culture, physical context etc.

UNIFICATION-BASED APPROACH FOR PARSING AND GENERATING SPEECH AND GESTURES

First studies at CST- formalisation and generation

- Unification-based approach – multimodal parser in HCI (i.a. deixis, feedback) (Paggio and Jongejan 2005)



hi nice sunshine today

```
Resulting Parse Tree # 0
Subgrammar[ 0 ]: Staging
I:[s_1adjunct,sem={act=comment,arg2={reln=sunshine,semtype=weather},advmod=today}]:{cat=s,stype=mai
 I:[reply_p_1]:{cat=reply_p}.[
  I:[hi_1]:{cat=adjunct,reln=hi,lex=hi,scat=no,semtype=none}.[
  ]
I:[s_np2,sem={act=comment,arg2={reln=sunshine,semtype=weather},advmod=today}]:{cat=s,stype=ellipsis}.[
I:[np_nodef2]:{cat=np,semtype=weather,gend=neut,nb=sing,defs=indef,pp_mod=no,mod=yes,coord=no,reln=
I:[ap_1]:{cat=ap,reln=nice,deg=base,sizegest=no,sizegestcont=no,scat=no,mod=yes,mod_n=yes,modsem=wea
 I:[nice_1]:{cat=adj,reln=nice,scat=no,deg=base,mod=yes,mod_n=yes,modsem=SEM,sizegest=no,lex=nice}.[
 I:[sunshine_1]:{cat=n,gend=neut,nb=sing,semtype=weather,reln=sunshine,lex=sunshine,scat=no,count=no}.[
   I:[ppadvp_1]:{cat=ppadvp,reln=today,semtype=SEM,scat=no}.[
   I:[advp_1]:{cat=advp,reln=today,semtype=SEM,scat=no}.[
    I:[today_1]:{cat=adv,reln=today,semtype=SEM,scat=no,lex=today}.[
```

comment_weather(sunshine,nice)

please take the cows to the well

```
Resulting Parse Tree # 0
Subgrammar[ 0 ]: Staging
I:[s_1adjunct,sem={act=request,predicate=take,arg2={reln=cow,semtype=animal,pointing=possible,pointobj=
 I:[reply_p_1]:{cat=reply_p}.[
  I:[please_1]:{cat=adjunct,reln=please,lex=please,scat=no,semtype=none}.[
I:[s_1imper,sem={act=request,predicate=take,arg2={reln=cow,semtype=animal,pointing=possible,pointobj=$r
I:[vp_arg12adv,sem={predicate=take,arg2={reln=cow,semtype=animal,pointing=possible,pointobj=$ref},advm
I:[take_3]:{cat=v,refl=no,scat=mainv,vform=infinitive,fin=no,voice=act,arg1=entity,arg2=entity,valency=arg12a
I:[np_def1]:{cat=np,pp_mod=no,semtype=animal,gend=neut,nb=plu,defs=def,mod=yes,coord=no,reln=cow,p
     I:[the_2]:{cat=det,gend=neut,nb=plu,defs=def,reln=the,lex=the,scat=no}.[
     I:[cows_1]:{cat=n,gend=neut,nb=plu,semtype=animal,reln=cow,lex=cows,scat=no,count=yes}.[
    I:[ppadvp_2]:{cat=ppadvp,reln=well,semtype=place}.[
I:[pp_1]:{cat=pp,reln=well,semtype=place,pform=to,ap_mod=MOD,ap_deg=DEG,sizegest=DEICT,sizegestco
      I:[to_1]:{cat=p,reln=to,lex=to,scat=no}.[
     I:[np_def1]:{cat=np,reln=well,semtype=place,ap_mod=MOD,ap_deg=DEG,sizegest=DEICT,sizegestcont
    I:[the_1]:{cat=det,gend=neut,nb=sing,defs=def,reln=the,lex=the,scat=no}.[
.I:[well_1]:{cat=n,gend=neut,nb=sing,semtype=place,reln=well,lex=well,scat=no,count=yes}.[
```

request(move(cow(all),well))

MUMIN annotation framework

(Allwood et al. 2007)

Attribute-value pairs to be used in an unification grammar describe:

- form
- semiotic type(s)
- function(s)
- emotion/attitude expression

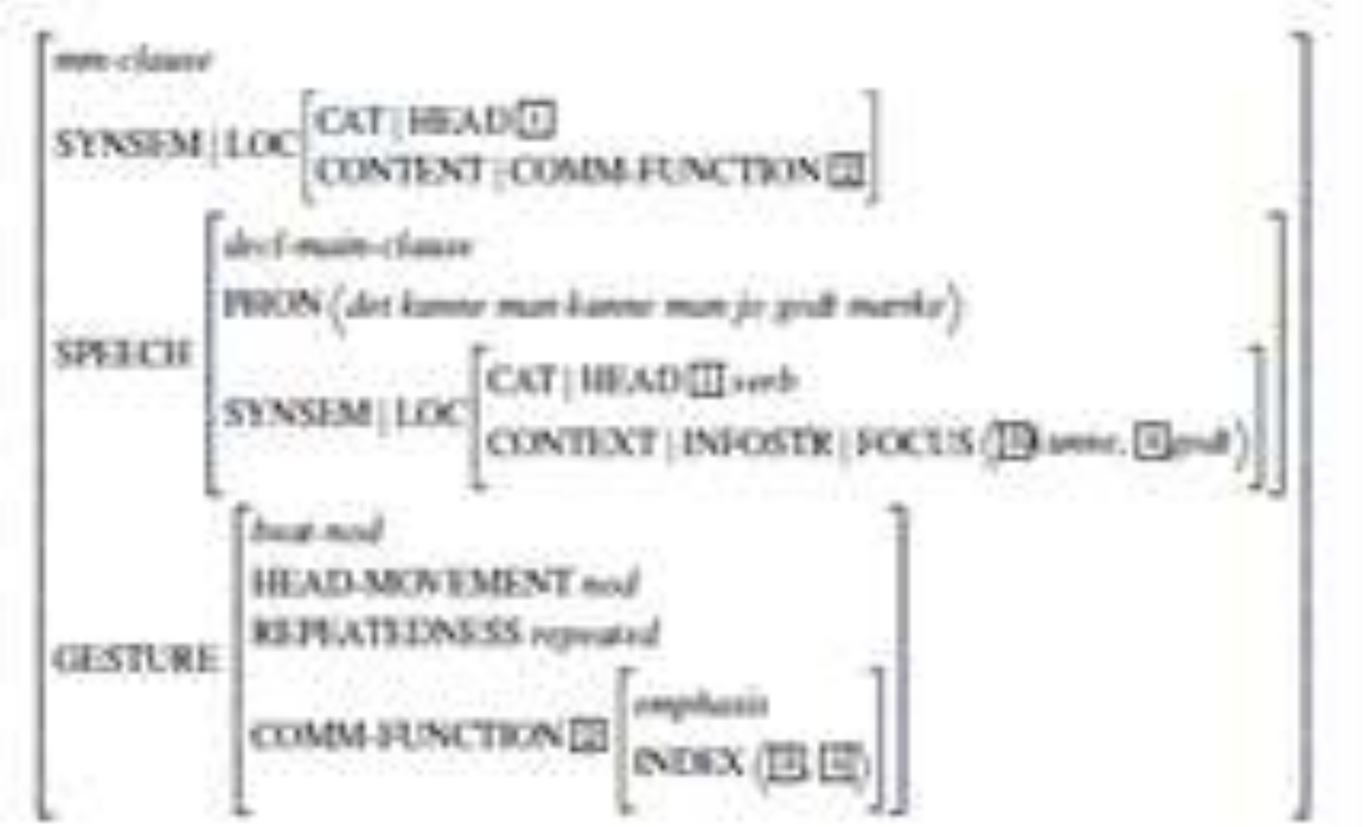
and links and relation type to connected speech/gestures

Example of feedback annotations

Attribute	Value
Feedback Basic	CPU (Contact + Perception+ Understanding), Other (C,CP)
Feedback Direction	Give, Elicit, GiveElicit, Underspecified
Feedback Agreement	Agree, Disagree

Formalising in HPSG

HPSG formalization of feedback and IS examples (Paggio and Navarretta 2009)



mm-clause

SYNSEM | LOC [CAT | HEAD $\boxed{1}$; CONTENT | COMM-FUNCTION [illegible]]

SPEECH [*decl-main-clause*

PHON ⟨*det kanne man kanne man jo godt mærke*⟩

SYNSEM | LOC [CAT | HEAD $\boxed{1}$ *verb* ; CONTEXT | INFOSTR | FOCUS ⟨[illegible] *kanne*, [illegible] *godt*⟩]]

GESTURE [*head-nod*

HEAD-MOVEMENT *nod*

REPEATEDNESS *repeated*

COMM-FUNCTION [illegible] [*emphasis* ; INDEX ⟨[illegible], [illegible]⟩]]

MULTIMODAL CORPUS COLLECTION AND ANNOTATION

NOMCO First encounters

- Comparable corpora: Danish, Finnish, Swedish

The Danish NOMCO first encounters

- Studio-recorded
- 6 females and 6 males: two first encounters each: 1F, 1M
- Participants stand in front of each other
- Three camera views
- Multiple multimodal annotations

(Paggio and Navarretta LRE 2017)

NOMCO project focus on Interaction Management (feedback and turn-taking)

Especially:

- Comparative studies on feedback signals
- Automatic identification of feedback and turn-taking gestures from gestures' shape and speech
- Temporal relation of facial expressions, head movements and speech
- Automatic identification of head movements from videos

Other studies on Danish NOMCO

- annotation of attitudes/emotions relevant to communication
  - relation of attitudes and communicative setting
  - relation of attitudes and speech
- transfer learning to annotate feedback in corpora in same language or in different languages
- automatic identification of gender and individuals from their communicative gestures

GESTURE TYPES

Semiotic types and meaning

Interpretation and semiotic types inspired by Peirce (1931):

- Indexical Non-deictic
  - Displays → Emotion, Feedback, Turn, Sequencing
  - Beats → Focusing, Feedback, Turn, Sequencing
- Indexical Deictic (pointing)→ Discourse referents, Turn
- Iconic
  - → representation of object/event
  - Metaphoric → representation of abstract idea/concept
- Emblem/symbol → Propositions.

Beat/Batonic and Deictic in the end

David Cameron in the Parliament 2009

https://www.youtube.com/watch?v=CBjnSsaIu70 , 0.53-1.06

Attitude expression, deictic, iconic reference to object

Meryl Streep Salutes Robert De Niro at the Kennedy Center Honors 2009

https://www.youtube.com/watch?v=coCxLpWUz50

0.50-1:10

Metaphoric gestures and iconic reference to event

https://www.youtube.com/watch?v=coCxLpWUz50

1:40-1:55

INTERPRETATION AND GENERATION OF HAND GESTURES

Relation between form and meaning (Kendon 2004)

(Navarretta LREC 2018)

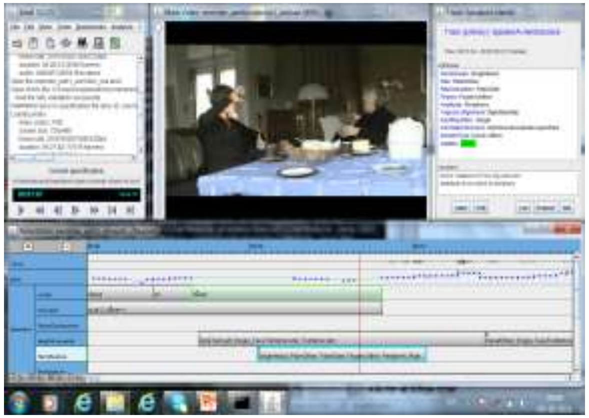
From the form of Obama's hand gestures predict their semiotic type (metaphoric gestures interpreted as simple iconic since speech not included)

Algorithm	P	R	F
Baseline	0.18	0.42	0.25
Bayes Network	0.59	0.6	0.59

Co-reference and gestures (Navarretta 2011)

Same pointing gesture co-occurs with co-referring expressions in map task corpus:. Gesture form contributes to resolution (Eisenstein and Davis 2006a, 2006b)

- similar iconic gestures should co-occur with co-referring expressions since their referent is the same (MOVIN/CLARIN-DK corpus)



Clustering using shape features of hand gestures

- Hand gestures co-occurring with co-referring expressions are grouped in the same cluster

- Only these exceptions (errors): speaker makes an iconic gesture co-occurring with an action and later points to the place where the iconic gesture was performed while referring to the action with abstract pronoun *det* (it/this/that)

Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs (Navarretta and **LIS** 2013)

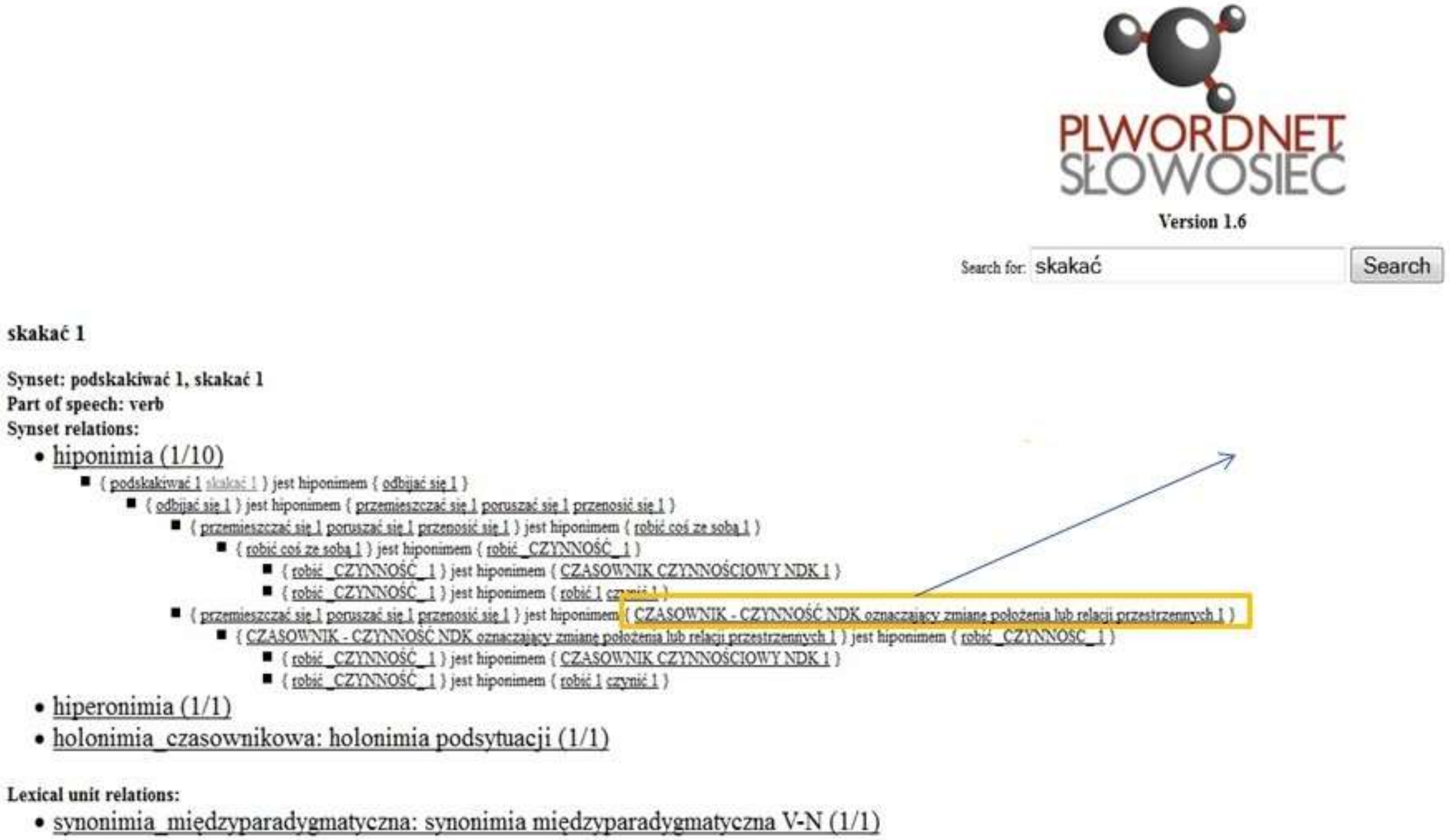
Gesture form

- **Viewpoint:** observer, character, (dual) (McNeill 1992)
- **Handedness**: which hand performs a gesture
- **Handshape**: the configuration of palm and fingers
- **Iteration**: static, single or repeated stroke
- **Movement**: the shape of the motion
- **Direction**: the plane on which it is performed

Factors influencing gesture form

- **Ontological type** of the referent (Poggi 2008) : Artefacts, Animates, Natural objects, Events
- **Events' structure** and viewpoint (Parrill 2010): Trajectory events - O-vpt, Handling events - C-vpt
- **Verb aspect** and handedness (Duncan 2002) and iteration (Parrill et al. 2013)
  - Perfective verbs - symmetric bi-handed
  - Imperfective - non-symmetric bi-handed
  - Progressive – iterated
- **Aktionsart** and speech-gesture temporal relationship (Becker et al. 2011)

(Lis 2012) Extends and operationalizes: Referent’s ontological type (Poggi 2008)+Events’ structure (Parrill 2010) using the Polish WordNet



skakać 1

Synset: podskakiwać 1, skakać 1
Part of speech: verb
Synset relations:

- hiponimia (1/10)
    - { podskakiwać 1 skakać 1 } jest hiponimem { odbijać się 1 }
        - { odbijać się 1 } jest hiponimem { przemieszczać się 1 poruszać się 1 przenosić się 1 }
            - { przemieszczać się 1 poruszać się 1 przenosić się 1 } jest hiponimem { robić coś ze sobą 1 }
                - { robić coś ze sobą 1 } jest hiponimem { robić _CZYNNOŚĆ_ 1 }
                    - { robić _CZYNNOŚĆ_ 1 } jest hiponimem { CZASOWNIK CZYNNOŚCIOWY NDK 1 }
                    - { robić _CZYNNOŚĆ_ 1 } jest hiponimem { robić 1 czynić 1 }
            - { przemieszczać się 1 poruszać się 1 przenosić się 1 } jest hiponimem { CZASOWNIK - CZYNNOŚĆ NDK oznaczający zmianę położenia lub relacji przestrzennych 1 }
                - { CZASOWNIK - CZYNNOŚĆ NDK oznaczający zmianę położenia lub relacji przestrzennych 1 } jest hiponimem { robić _CZYNNOSC_ 1 }
                    - { robić _CZYNNOŚĆ_ 1 } jest hiponimem { CZASOWNIK CZYNNOŚCIOWY NDK 1 }
                    - { robić _CZYNNOŚĆ_ 1 } jest hiponimem { robić 1 czynić 1 }
- hiperonimia (1/1)
- holonimia_czasownikowa: holonimia podsytuacji (1/1)

Lexical unit relations:

- synonimia_międzyparadygmatyczna: synonimia międzyparadygmatyczna V-N (1/1)

Speech annotation

ATTRIBUTE	VALUE
Event subtype	Translocation_spatial_relations Non-translational motion Other
Aktionsart	State, Act, Activity, Accident, Event, Action, Process
Aspect	Perfective, Imperfective

Gesture annotation

ATTRIBUTE	VALUE
Viewpoint	C-vpt, O-vpt, D-vpt
Handedness	Right_Hand, Left_Hand, 2 _Symmetric _Hands, 2_NonSymmetric_Hands
Handshape	Claw, Fist, IndexFinger, Open, Purse, Ring, Other
Iteration	Single_Stroke, Repeated_Stroke, Stroke_Hold
Movement	Straight, Arc, Circle, Complex, None
Direction	Horizontal_Saggital, Horizontal_Coronal, Vertical, Multidirectional, None

The data

- audio and video-recordings
- 5 male, 5 female Polish speakers
- retelling task (McNeill 1992, 2005)

Classification of Handshape

(support vector machine)

Handshape	Precision	Recall	F-score
baseline	0.08	0.28	0.12
aspect	0.08	0.28	0.12
aktionsart	0.26	0.33	0.26
verb form	0.28	0.34	0.26
(sub)type	**0.28**	**0.36**	**0.29**
all	**0.38**	**0.37**	**0.35**

Classification of Handedness

Handedness	Precision	Recall	F-score
baseline	0.19	0.44	0.27
aspect	0.19	0.44	0.27
aktionsart	0.3	0.4	0.34
verb form	**0.44**	**0.53**	**0.45**
(sub)type	0.32	0.44	0.37
all	**0.45**	**0.49**	**0.47**

Classification of Direction

Direction	Precision	Recall	F-score
baseline	0.13	0.36	0.19
aspect	0.27	0.38	0.31
aktionsart	0.27	0.38	0.31
verb form	**0.41**	**0.41**	**0.41**
(sub)type	0.29	0.36	0.29
all	**0.45**	**0.45**	**0.44**

Predicting the Viewpoint from the linguistic information

Viewpoint	Precision	Recall	F-score
baseline	0.29	0.54	0.38
aspect	0.29	0.54	0.38
aktionsart	0.54	0.60	0.54
verb form	0.69	0.68	0.65
(sub)type	**0.70**	**0.78**	**0.73**
all	**0.78**	**0.79**	**0.77**

Predicting the Viewpoint from gesture’s form

Viewpoint	Precision	Recall	F-score
baseline	0.29	0.54	0.38
handshape	**0.58**	**0.65**	**0.61**
handness	**0.54**	**0.61**	**0.57**
iteration	0.59	0.56	0.42
movement	0.65	0.56	0.43
direction	0.48	0.56	0.49
all	**0.67**	**0.68**	**0.67**

Conclusions

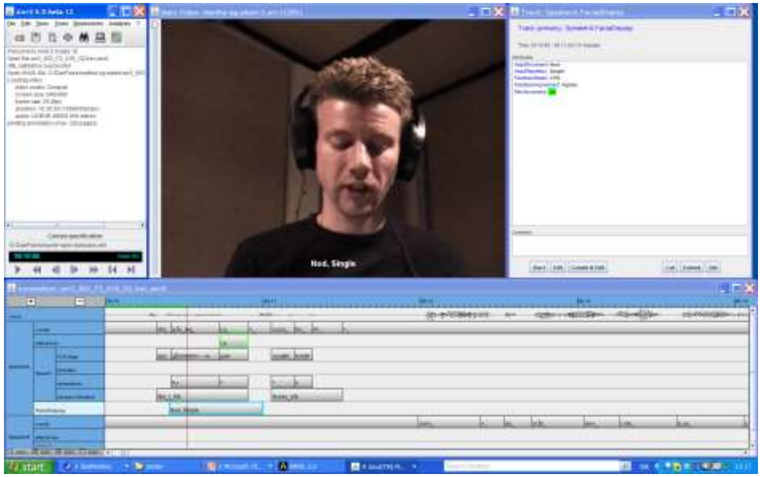
Even if there are many individual differences in the way people perform iconic gestures referring to the same event, the relation between the semantic type of the referent and some form/shape features of the referring hand gestures holds in the same language. Interesting for gesture generation.

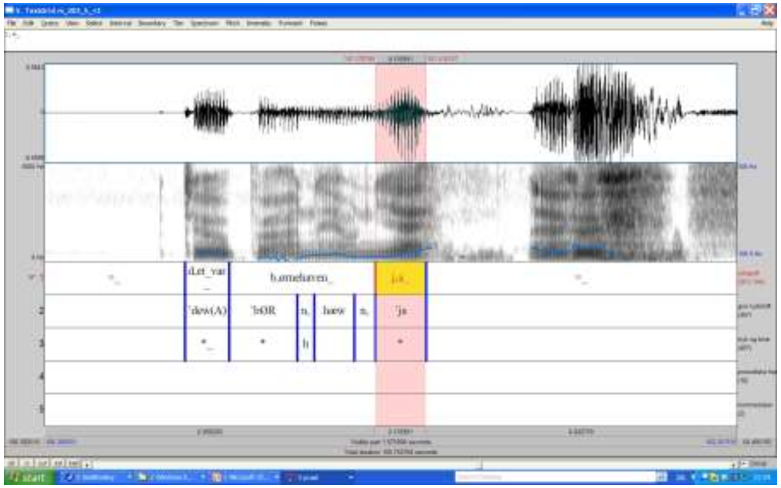
PROSODY, GESTURES AND DIALOGUE ACTS RELATED TO FEEDBACK WORDS

Aims

(Navarretta and Paggio 2010)

- Can prosodic features, head movements and facial expressions disambiguate the meaning (dialogue acts) of Yes and No expressions: *ja* (yes), *jo* (yes in a negative context), *jamen* (yes but, well), *nej* (no), *næh* (no) in Maptask interactions.





Annotations

Danish DanPASS corpus (Grønnum 2006):

- Transcriptions and prosodic features: stress, tone and hesitations

Our annotations:

- Dialogue Acts for *Yes/No* expressions (subset of

ISO 24617-2): Agreement, Disagreement, Answer, Repeat-Rephrase, Accept

- Facial expressions and head movements: subset of MUMIN categories.

Classification 1: dialogue acts and prosody

820 Yes and 96 No expressions without gestures
Hidden Naïve Bayes, baseline Majority classifier

Dataset	P	R	F
YesNo	*27.8*	*52.8*	*36.5*
YesNo	47.2	53	46.4
+stress	47.5	54.1	47.1
+stress+tone	**47.8**	**54.3**	**47.4**
+stress+tone+ hesitation	47.7	54.5	47.3

Classification 2: prosody + gestures

- Yes and No expressions accompanied by gesture (204 Yes, 24 No):

Dataset	Algorithm	P	R	F
YesNo+stress+tone	*HNB*	*43.1*	*56.1*	*46.4*
+face	HNB	43.7	56.1	46.9
+headmovement	HNB	47	57.9	51
+face+headmovem	HNB	**51.6**	**57.9**	**53.9**

TEST INFLUENCE OF SPEECH PAUSES AND GESTURES ON SUCCESSFUL RESPONSE FROM AUDIENCE

Prediction of Audience Response in Humorous Discourse by Barack Obama

(Navarretta 2017)

What? **pauses (***silent* or *filled*, e.g. *um*, *ah*, *uh*) and **gestures**: head movements, facial expressions and hand gestures.

Why?
Pauses and gestures have multiple and often co-occurring functions and their importance in (humorous) speech has been addressed in various studies.

Aims

- Determine to what extent information about sequences of audience response, spoken segments, speech pauses and co-occurring gestures can predict the success of humorous talks. Success is measured as immediate audience response (laughter and/or cheers and/or applause).

Pauses

Are voluntary or involuntary signals that:

- regulate the interaction (Duncan and Fiske 1977,Clark and Fox Tree 2002)
- indicate that speakers are:

  a) planning and structuring the message (Maclay and Osgood 1959, Chafe 1987)

  b) looking for the appropriate word (Rochester 1973) or going to present difficult/abstract concepts (Reynolds and Paivio 1968), complex anaphora (Navarretta 2010)

Pauses in humorous speech and comedy

- Structure and emphasize the discourse, give time to reflect on conveyed message (Sankey 1998, Oliver 2013). Speech rate the same in humorous and non-humorous discourse, pauses do not precede punch lines (Attardo et al. 2011), but speakers laugh more when presenting humorous speech (Attardo and Pickering 2011)

Pauses in political speech

- Silent pauses are 50% longer in televised political speeches than in general or political interviews (Duez 1982). Emphatic pauses are more frequent in political speeches than in other speech types held by Italian Silvio Berlusconi (Salvati and Pettorini 2010).

Gesture and pauses

- Gestures and pauses are temporal and functional related i.a. (Kendon 1964, 1967, Dittman 1972, Butterworth and Hadar 1989, Esposito et al. 2001, Esposito and Esposito 2011).

- Obama: excellent speaker, great presentation style (Cooper 2011).

Audience response

- Guerini et al. (2010) add occurrences of audience reaction (applause, laughter, other) to transcriptions of American political speeches in order to find prominent discourse segments.

The data

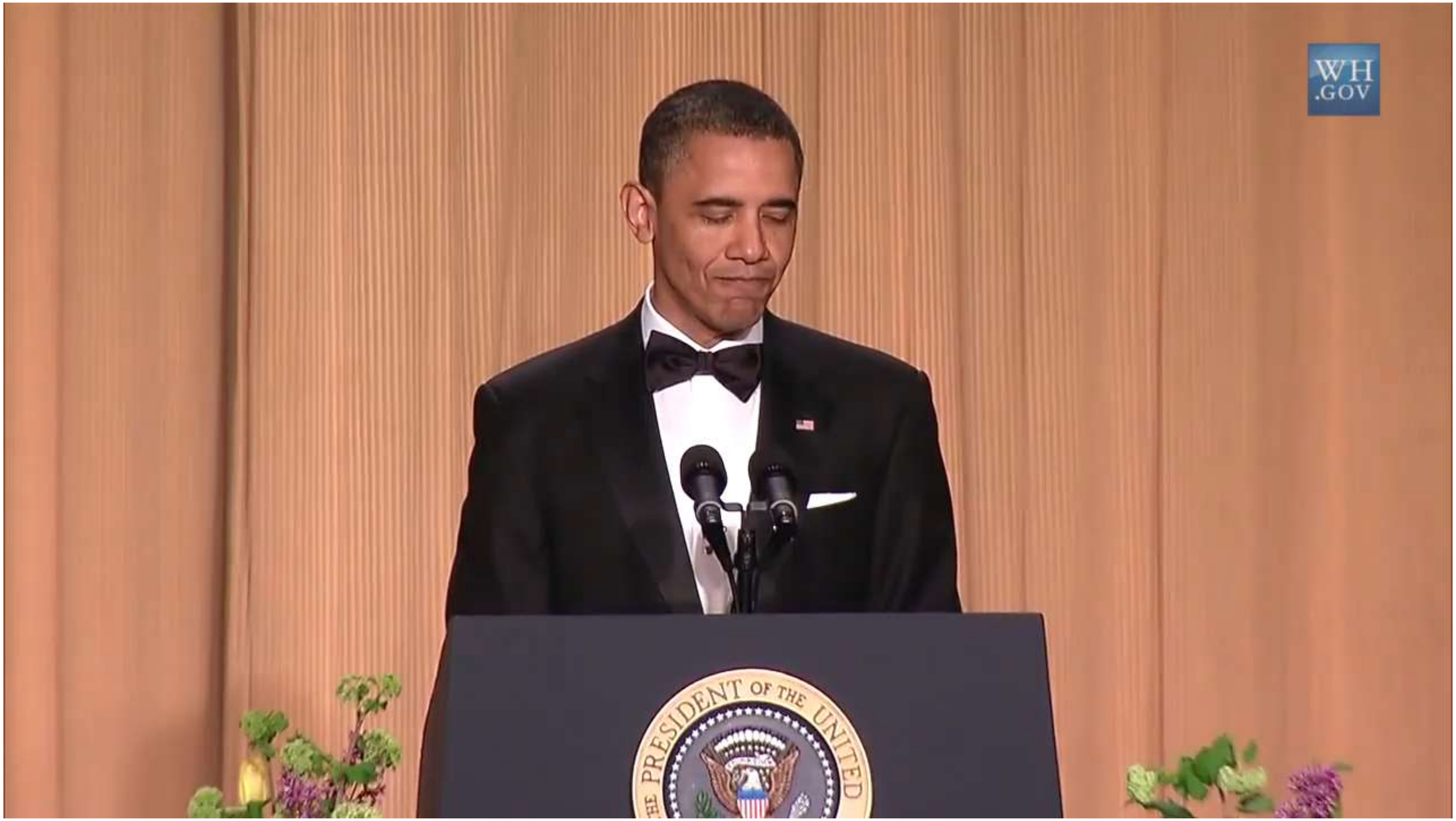
Speeches by Barack Obama at the Annual White House Correspondents' Association Dinner in 2011 and 2016. Videos were downloaded from http:\\www.WH.gov

Speech parts:

- talk11: 13 min and 22 sec
- talk16: 30 minutes

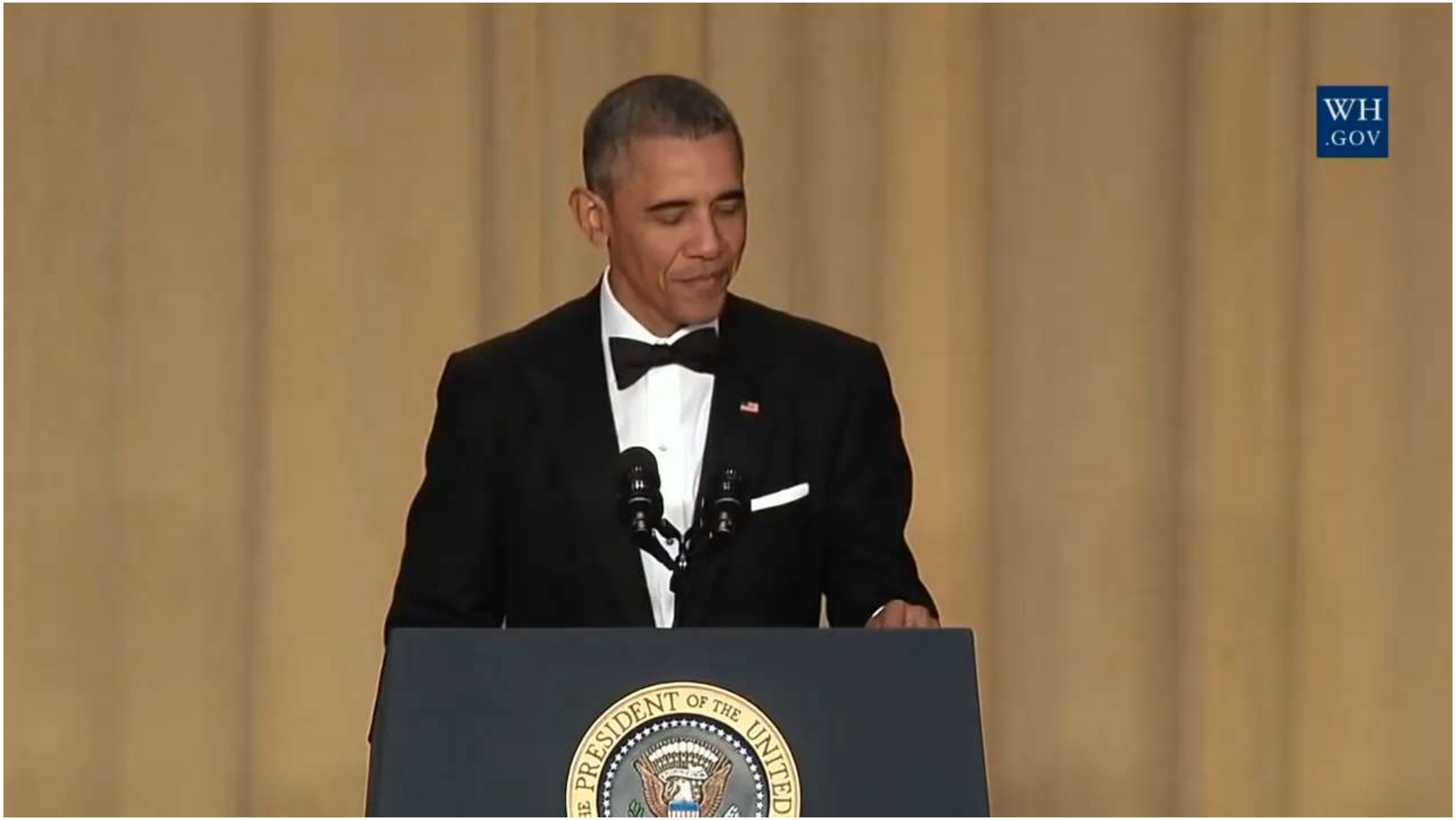
2011

whole speech can be seen at
https://www.youtube.com/watch?v=n9mzJhvC-8E



2016

whole speech can be seen at
https://www.youtube.com/watch?v=TO9d16c2XRM



The annotations

- Speech pauses longer than 0.06 sec. extracted automatically in PRAAT.
- Speech sequences were inserted between pauses semi-automatically using text taken from official transcriptions.
- Gestures were manually annotated according to MUMIN scheme (Allwood et al. 1997)

Gestural features

Attribute	Value
HeadMovement	Nod, UpNod, HeadForward, Tilt, HeadBackward, SideTurn, Shake, Waggle, HeadOther
HeadRepeat	HeadSingle, HeadRepeated
Face	Smile, Laugh, Scowl, FaceOther, EyebrowRaise, Frown, EyebrowsOther
Handedness	BothHandsSym, BothHandsAsym, RightSingleHand, LeftSingleHand
HandRepeat	HandSingle, HandRepeated

Data analysis

- Pauses are used voluntary to emphasize and structure the discourse.
- Same speech rate, and same relative frequency of head movements and facial expressions in 2011 and 2016
- Significantly more hand gestures in 2016 than in 2011 (χ square = 19.295 with 1 df, and 2-tailed p<0.0001).
- **Audience response and pauses positively correlated: Pearson 2-tailed correlation r=0.465 and r(1541)<0.0001**.

Predicting audience response

- Neural network trained on unigrams, bigrams, trigrams consisting of:
- sequences of speech segments, pauses, external contexts (videos, music etc.) and audience reaction

  + information of duration (all tokens)

  + information of co-occurring gestures by Obama

Multimodal unit

Duration	Speech	Face	Head	Hand
0.78	spoken-seq	B-raise	nod-single	none
0.56	pause	none	none	none
1.99	spoken-seq	none	none	both-hand-sin
3.64	audience	none	forward-single	none
0.85	spoken-seq	smile	none	none
0.1	pause	smile	none	none

Trigrams of multimodal units

multimodal unit	multimodal unit	multimodal unit	Response
0.78, spoken-seq, B-raise, node single, none	0.56, pause, none, none, none	1,99, spoken-seq, none, none, both hands	Yes
0.56, pause, none, none, none	1,99, spoken-seq, none, none, both hands	3.64, audience, none, forward-single, none, none	No

Data	P	R	F-score
Majority classifier	0.69	0.82	0.76
Unigrams: audio	0.72	0.83	0.77
Unigrams: audio + duration	0.69	0.83	0.76
Unigrams: audio + gestures	0.69	0.83	0.76
Bigrams: audio	0.69	0.82	0.76
Bigrams: audio + duration	0.81	0.84	0.81
Bigrams: audio + gestures	0.86	0.88	0.86
Bigrams: audio + duration + gesture	0.87	0.88	0.87
Trigrams: audio	0.82	0.83	0.83
Trigrams: audio + duration	0.82	0.85	0.84
Trigrams: audio + gesture	0.87	0.88	0.87
Trigrams: audio + duration + gesture	**0.88**	**0.89**	**0.88**

Discussion

- Results confirm that speech pauses and gestures are important for presenting (humorous) message successfully and information about them can contribute to systems for training humans, talking software agents and robots.

However,

- content is the must important element in (humorous) speech and it is not addressed in this work.

Future

- Is it possible to predict both positive and negative audience response in these speeches and in other speech types?

Concluding

All these studies confirm that speech and gestures are strongly related in both the production and reception of face-to-face communication.

The analysed data is still too small: there are few freely available annotated corpora.

The automatic identification and interpretation of gestures from existing videos is still not good, but sensors and wearable devises can be used to analyse new data.

QUESTIONS?