Do we need explicit models of prosodic form to interpret spoken data?

-

Workshop OTIM/ILIKS
LPL, Aix-en-Provence

Daniel Hirst

Laboratoire Parole et Langage, CNRS and Université de Provence
daniel.hirst@lpl-aix.fr

2011-05-24

The curse of Babel?

Figure:

The curse of Babel

- The language barrier is perhaps the greatest social problem facing modern multicultural societies like Europe.

The curse of Babel

- The language barrier is perhaps the greatest social problem facing modern multicultural societies like Europe.
- Language is not just words - non-verbal information is (at least) just as important.

The curse of Babel

- The language barrier is perhaps the greatest social problem facing modern multicultural societies like Europe.
- Language is not just words - non-verbal information is (at least) just as important.
- This is an area where we need speech technology.

The curse of Babel

- The language barrier is perhaps the greatest social problem facing modern multicultural societies like Europe.
- Language is not just words - non-verbal information is (at least) just as important.
- This is an area where we need speech technology.
- Speech technology for non-verbal information is in its infancy.

What is missing?

What is missing?

Figure: Why can't we use these to speak to people in other languages?

What have we already got?

What have we already got?

- Speech recognition (Dragon dictate, Google translate)

What have we already got?

- Speech recognition (Dragon dictate, Google translate)
- Translation (Babelfish, Google translate)

What have we already got?

- Speech recognition (Dragon dictate, Google translate)
- Translation (Babelfish, Google translate)
- Speech synthesis (Acapela, Google translate)

What have we already got?

- Speech recognition (Dragon dictate, Google translate)
- Translation (Babelfish, Google translate)
- Speech synthesis (Acapela, Google translate)

What have we already got?

- Speech recognition (Dragon dictate, Google translate)
- Translation (Babelfish, Google translate)
- Speech synthesis (Acapela, Google translate)

Figure: My hovercraft is full of eels!

Speech technology

- current disparity in resources

Speech technology

- current disparity in resources
- small minority of languages - acceptable (?)

Speech technology

- current disparity in resources
- small minority of languages - acceptable (?)
- vast majority of languages - primitive

Speech technology

- current disparity in resources
- small minority of languages - acceptable (?)
- vast majority of languages - primitive
- transfer of ressources?

Speech technology resources

- often language specific

Speech technology resources

- often language specific
- difficult to generalise to:

Speech technology resources

- often language specific
- difficult to generalise to:
- - under-ressourced languages

Speech technology resources

- often language specific
- difficult to generalise to:
- - under-ressourced languages
- - different dialects

Speech technology resources

- often language specific
- difficult to generalise to:
- - under-ressourced languages
- - different dialects
- - different speaking styles

Speech technology resources

- often language specific
- difficult to generalise to:
- - under-ressourced languages
- - different dialects
- - different speaking styles
- speech prosody

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"
- statement? question? order?

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"
- statement? question? order?
- speaker states "This is really interesting"

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"
- statement? question? order?
- speaker states "This is really interesting"
- naturalness

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"
- statement? question? order?
- speaker states "This is really interesting"
- naturalness
- - facilitate cognitive processing

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"
- statement? question? order?
- speaker states "This is really interesting"
- naturalness
- - facilitate cognitive processing
- - cf non-standard, non-native, pathological, or synthetic speech

Annotation of speech prosody

The annotation/representation of prosody is crucial for

- intelligibility "He’s not coming back"
- statement? question? order?
- speaker states "This is really interesting"
- naturalness
- - facilitate cognitive processing
- - cf non-standard, non-native, pathological, or synthetic speech
- limited current use of synthesis for listening tasks but huge potential

Annotation of speech prosody

Current prosodic annotation is too language / theory specific

- cross-language annotation

Annotation of speech prosody

Current prosodic annotation is too language / theory specific

- cross-language annotation
- - INTSINT (Hirst & Di Cristo 1998)

Annotation of speech prosody

Current prosodic annotation is too language / theory specific

- cross-language annotation
- - INTSINT (Hirst & Di Cristo 1998)
- - ToBI (Jun 2005)

Annotation of speech prosody

Current prosodic annotation is too language / theory specific

- cross-language annotation
- - INTSINT (Hirst & Di Cristo 1998)
- - ToBI (Jun 2005)
- interaction between linguists and engineers

Annotation of speech prosody

Current prosodic annotation is too language / theory specific

- cross-language annotation
- - INTSINT (Hirst & Di Cristo 1998)
- - ToBI (Jun 2005)
- interaction between linguists and engineers
- Biannual Speech Prosody Conferences

Annotation of speech prosody

Current prosodic annotation is too language / theory specific

- cross-language annotation
- - INTSINT (Hirst & Di Cristo 1998)
- - ToBI (Jun 2005)
- interaction between linguists and engineers
- Biannual Speech Prosody Conferences
- 6th International Speech Prosody Conference,
  (May 2012 - Shanghai)

Prosodic annotation function vs form

- most prosodic annotation systems don't distinguish

Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%

Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
- function (* %)

Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
- function (* %)
- form (HL)

Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
- function (* %)
- form (HL)
- Inter-transcriber agreement (Wightman 2002)

Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
- function (* %)
- form (HL)
- Inter-transcriber agreement (Wightman 2002)
- - functions good

Prosodic annotation function vs form

- most prosodic annotation systems don't distinguish
- ToBI: H* L%
- function (* %)
- form (HL)
- Inter-transcriber agreement (Wightman 2002)
- - functions good
- - forms bad

Prosodic annotation function vs form

- most prosodic annotation systems don't distinguish
- ToBI: H* L%
- function (* %)
- form (HL)
- Inter-transcriber agreement (Wightman 2002)
- - functions good
- - forms bad
- Automatic recognition the opposite

Prosodic form

- Momel/INTSINT

Prosodic form

- Momel/INTSINT
- Automatic reversible annotation with Momel

Prosodic form

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into

Prosodic form

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
- - macroprosodic component
  (independent of segmental material)

Prosodic form

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
- - macroprosodic component
  (independent of segmental material)
- - microprosodic component
  (independent of intonation)

Prosodic form

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
- - macroprosodic component
  (independent of segmental material)
- - microprosodic component
  (independent of intonation)
- INTSINT designed as tool for linguists
  for the symbolic coding of intonation patterns.
  (Hirst & Di Cristo (eds) 1998)

Prosodic form

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
- - macroprosodic component
  (independent of segmental material)
- - microprosodic component
  (independent of intonation)
- INTSINT designed as tool for linguists
  for the symbolic coding of intonation patterns.
  (Hirst & Di Cristo (eds) 1998)
- Both now implemented as plugin for Praat

INTSINT to Momel

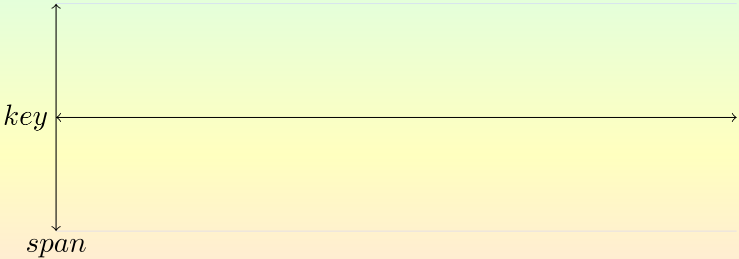


Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

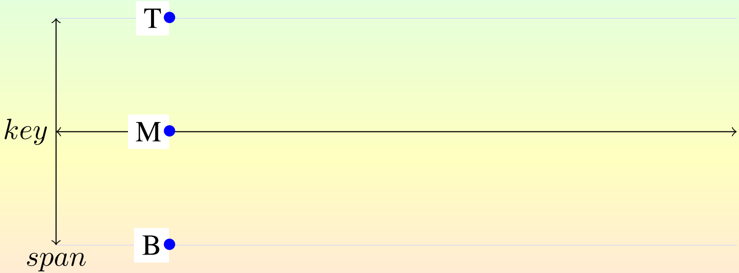


Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

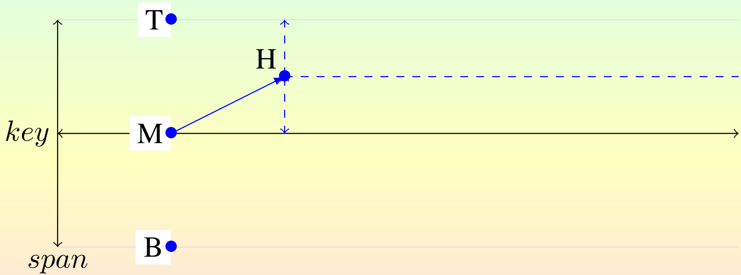


Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

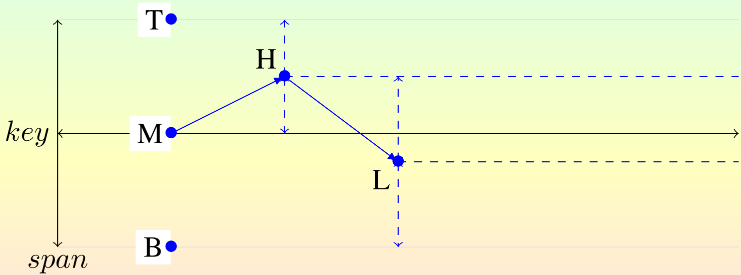


Figure: INTSINT to MoMel defined by 2 parameters *key* and *span*

INTSINT to Momel

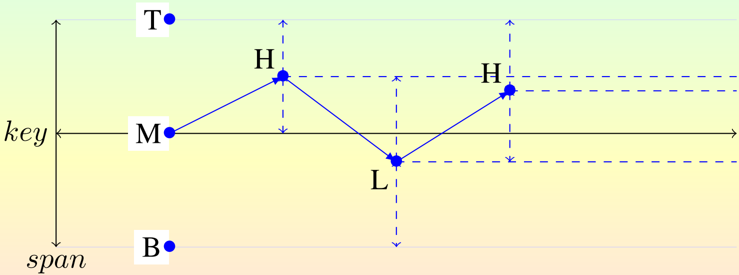


Figure: INTSINT to MoMel defined by 2 parameters *key* and *span*

INTSINT to Momel

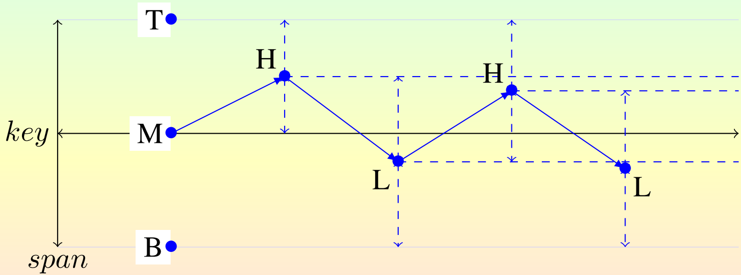


Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

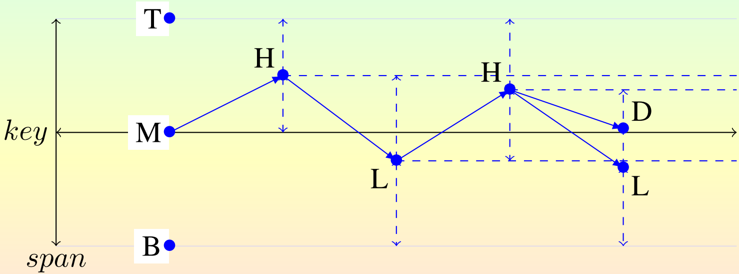


Figure: INTSINT to MoMel defined by 2 parameters *key* and *span*

Prosodic function

- IF annotation (Hirst 1977, 2005)

Prosodic function

- IF annotation (Hirst 1977, 2005)
- 4 degrees of prominence
  unaccented, accented, nuclear, emphatic

Prosodic function

- IF annotation (Hirst 1977, 2005)
- 4 degrees of prominence
  unaccented, accented, nuclear, emphatic
- 3 degrees of boundary
  none, non-terminal, terminal

Prosodic function

- IF annotation (Hirst 1977, 2005)
- 4 degrees of prominence
  unaccented, accented, nuclear, emphatic
- 3 degrees of boundary
  none, non-terminal, terminal
- label a large and sufficiently representative corpus:
  in terms of the higher-level factors that govern phonemic, phrasal, prosodic, speech-act etc. variation. (Campbell 1995)

Bootstrapping automatic prosodic functional annotation

- Hand-labelled data on small corpus

Bootstrapping automatic prosodic functional annotation

- Hand-labelled data on small corpus
- Predict functional annotation from acoustic data

Bootstrapping automatic prosodic functional annotation

- Hand-labelled data on small corpus
- Predict functional annotation from acoustic data
- Train synthesiser with larger corpus of annotated data

Application to TTS in Finnish

Vainio, Hirst, Suni & De Looze (in Proc. SpeCom 2009)

- HMM based system

Application to TTS in Finnish

Vainio, Hirst, Suni & De Looze (in Proc. SpeCom 2009)

- HMM based system
- symbolic input sequence of phone-sized HMM units

Application to TTS in Finnish

Vainio, Hirst, Suni & De Looze (in Proc. SpeCom 2009)

- HMM based system
- symbolic input sequence of phone-sized HMM units
- prosodic parameters: F0, duration, glottal flow

Application to TTS in Finnish

Vainio, Hirst, Suni & De Looze (in Proc. SpeCom 2009)

- HMM based system
- symbolic input sequence of phone-sized HMM units
- prosodic parameters: F0, duration, glottal flow
- training data not labelled for prosodic form

Application to TTS in Finnish

Vainio, Hirst, Suni & De Looze (in Proc. SpeCom 2009)

- HMM based system
- symbolic input sequence of phone-sized HMM units
- prosodic parameters: F0, duration, glottal flow
- training data not labelled for prosodic form
- iterative procedure: train on functional annotation

Application to TTS in Finnish

Vainio, Hirst, Suni & De Looze (in Proc. SpeCom 2009)

- HMM based system
- symbolic input sequence of phone-sized HMM units
- prosodic parameters: F0, duration, glottal flow
- training data not labelled for prosodic form
- iterative procedure: train on functional annotation
- predict prosodic tags from hand-labelled corpus

Sample synthesis

(using functional annotation)

Viron Pärnussa vesi on yön aikana vetäytynyt pääosin takaisin merelle. Pelastustyöt kuitenkin jatkuvat, eikä evakuoituja ihmisiä voida Viron television mukaan todennäköisesti siirtää takaisin en nen iltaa. Ensin tarkistetaan, ovatko talot kunnossa. Haapsalun suunnalla evakuoitujen ihmisten on luvattu pääsevän takaisin jo aiemmin. Sääennusteen mukaan tänään voi Virossa sataa ja tuulla kovaa.

Application to synthesis of French

- Read speech: corpus Eurom1 (-> Multext Prosody):

Application to synthesis of French

- Read speech: corpus Eurom1 (-> Multext Prosody):
- - 40 continuous passages of 5 sentences each.

Application to synthesis of French

- Read speech: corpus Eurom1 (-> Multext Prosody):
- - 40 continuous passages of 5 sentences each.
- Spontaneous speech: corpus CID (Bertrand et al. 2008):

Application to synthesis of French

- Read speech: corpus Eurom1 (-> Multext Prosody):
- - 40 continuous passages of 5 sentences each.
- Spontaneous speech: corpus CID (Bertrand et al. 2008):
- - interactive dialogue: 8 one-hour dialogues.

Application to synthesis of French

- Read speech: corpus Eurom1 (-> Multext Prosody):
- - 40 continuous passages of 5 sentences each.
- Spontaneous speech: corpus CID (Bertrand et al. 2008):
- - interactive dialogue: 8 one-hour dialogues.
- Each dialogue about 20 minutes for each speaker.

Application to synthesis of French

- Read speech: corpus Eurom1 (-> Multext Prosody):
- - 40 continuous passages of 5 sentences each.
- Spontaneous speech: corpus CID (Bertrand et al. 2008):
- - interactive dialogue: 8 one-hour dialogues.
- Each dialogue about 20 minutes for each speaker.
- Treat each speech style as different language

So no future for explicit models of prosodic form?

- not for labelling but for evaluation

So no future for explicit models of prosodic form?

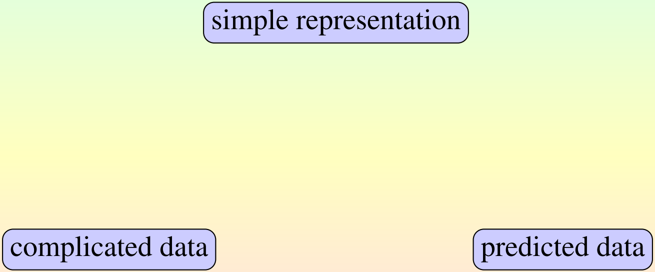
- not for labelling but for evaluation
- analysis by synthesis

So no future for explicit models of prosodic form?

- not for labelling but for evaluation
- analysis by synthesis

Analysis by synthesis

Analysis by synthesis



simple representation

complicated data

predicted data

Figure: The Analysis by Synthesis paradigm

Analysis by synthesis

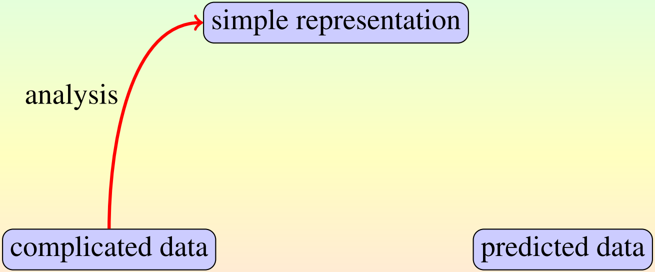


Figure: The Analysis by Synthesis paradigm

Analysis by synthesis

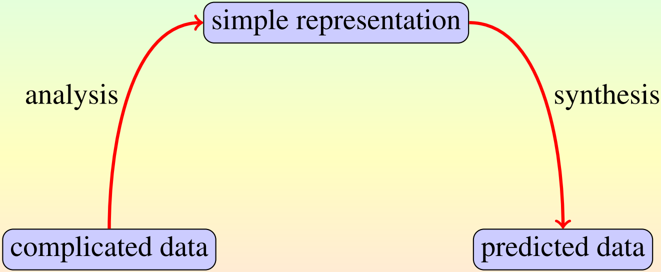


Figure: The Analysis by Synthesis paradigm

Analysis by synthesis

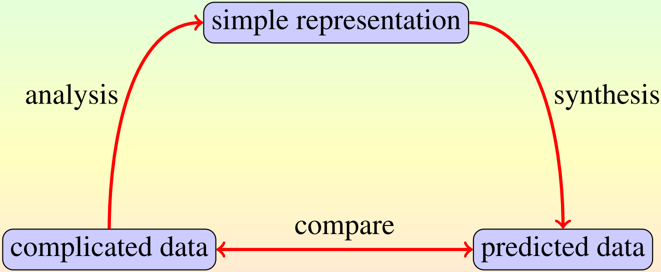


Figure: The Analysis by Synthesis paradigm

What is science?

What is science?

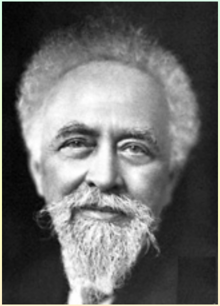


Figure: Jean Baptiste Perrin (1870-1942).

scientific method: explain visible complexity
by invisible simplicity.
(expliquer le visible compliqué par l’invisible simple.)