The analysis by synthesis of speech prosody.

Speech Lunch, Phonetics Laboratory
University of Oxford

Daniel Hirst

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

2011-10-14

The curse of Babel?

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?

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

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
- small minority of languages - acceptable (?)
- vast majority of languages - primitive
- transfer of ressources?

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"
- statement? question? order?
- speaker states "Isn't this 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
  - - 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 "ToBI or not ToBI")

Prosodic annotation function vs form

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

Phonetic representation

- Momel/INTSINT

Phonetic representation

- Momel/INTSINT
- Automatic reversible annotation with Momel

Phonetic representation

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

Phonetic representation

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

Phonetic representation

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

Momel

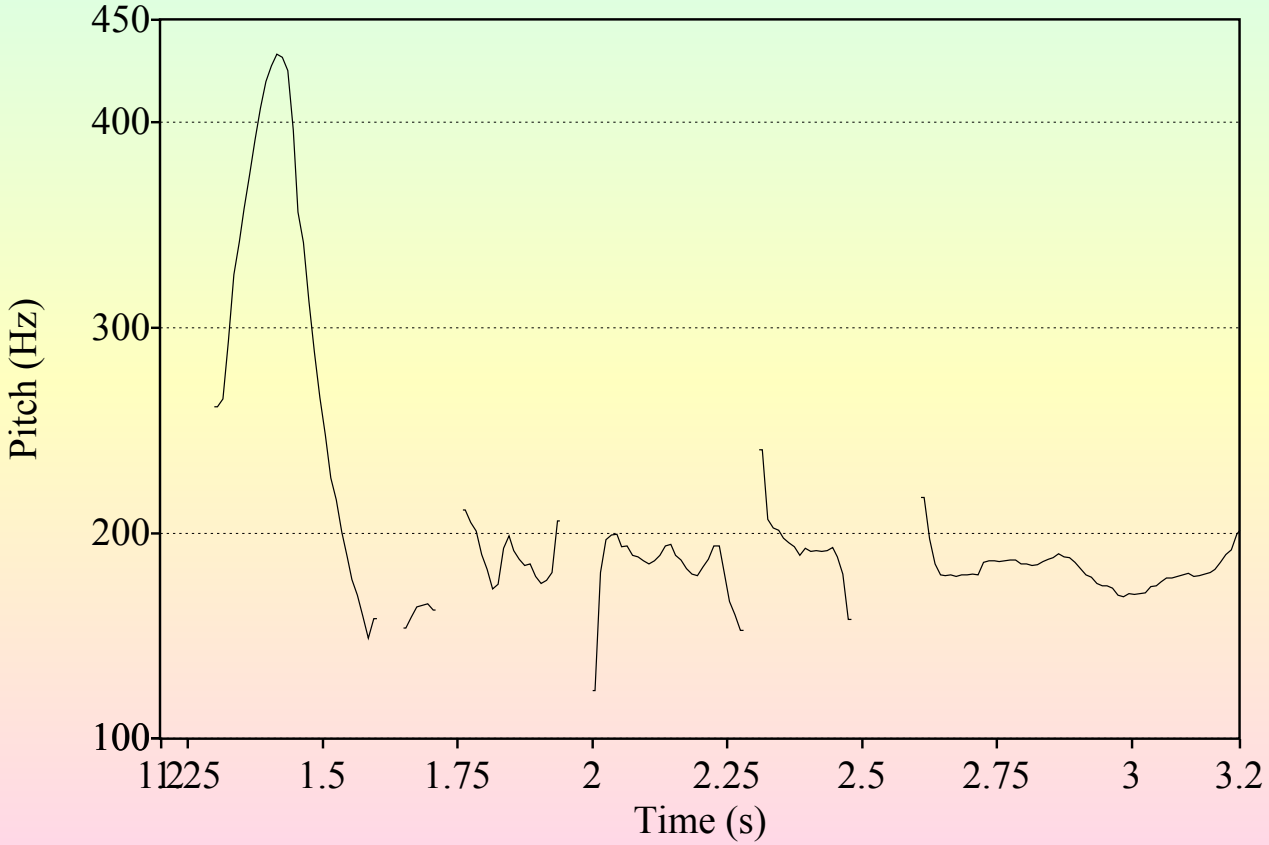

Figure: Momel

Momel

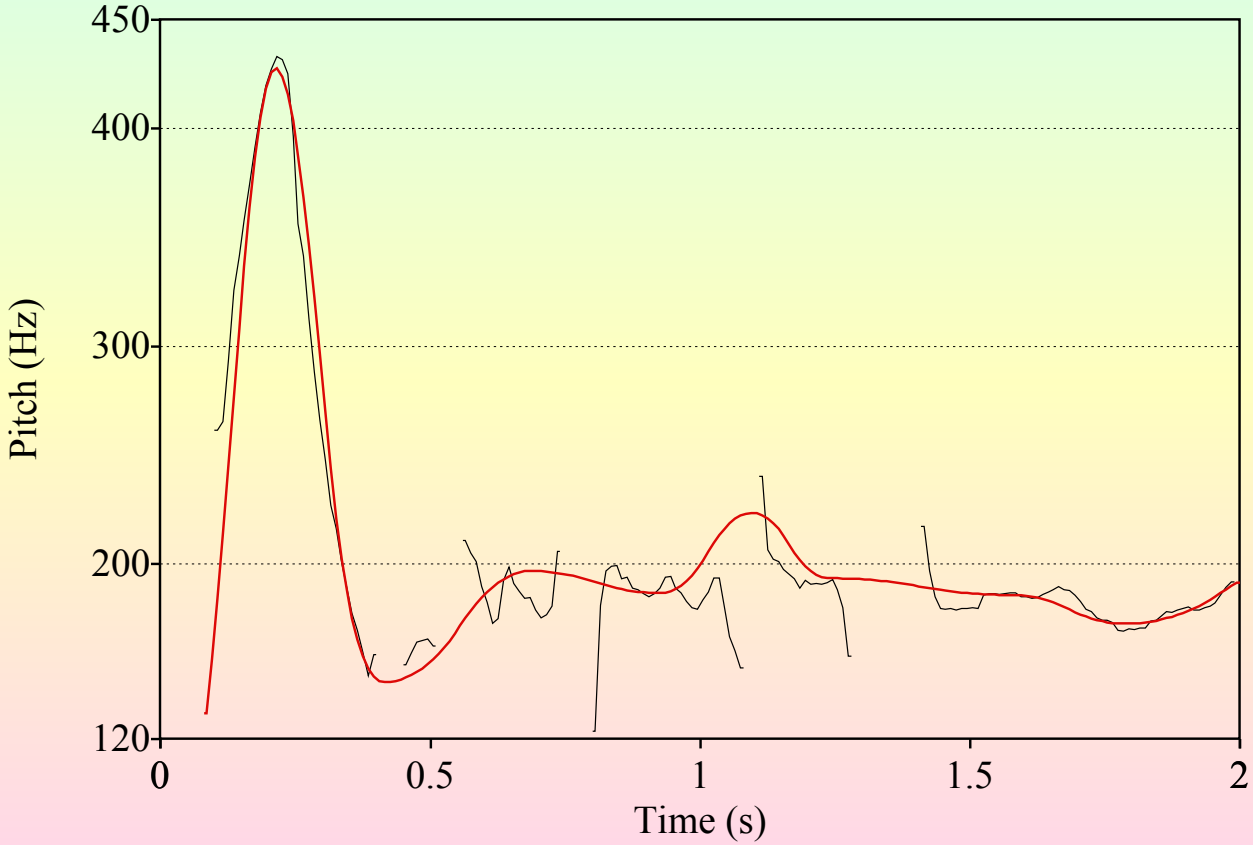

Figure: Momel

Momel

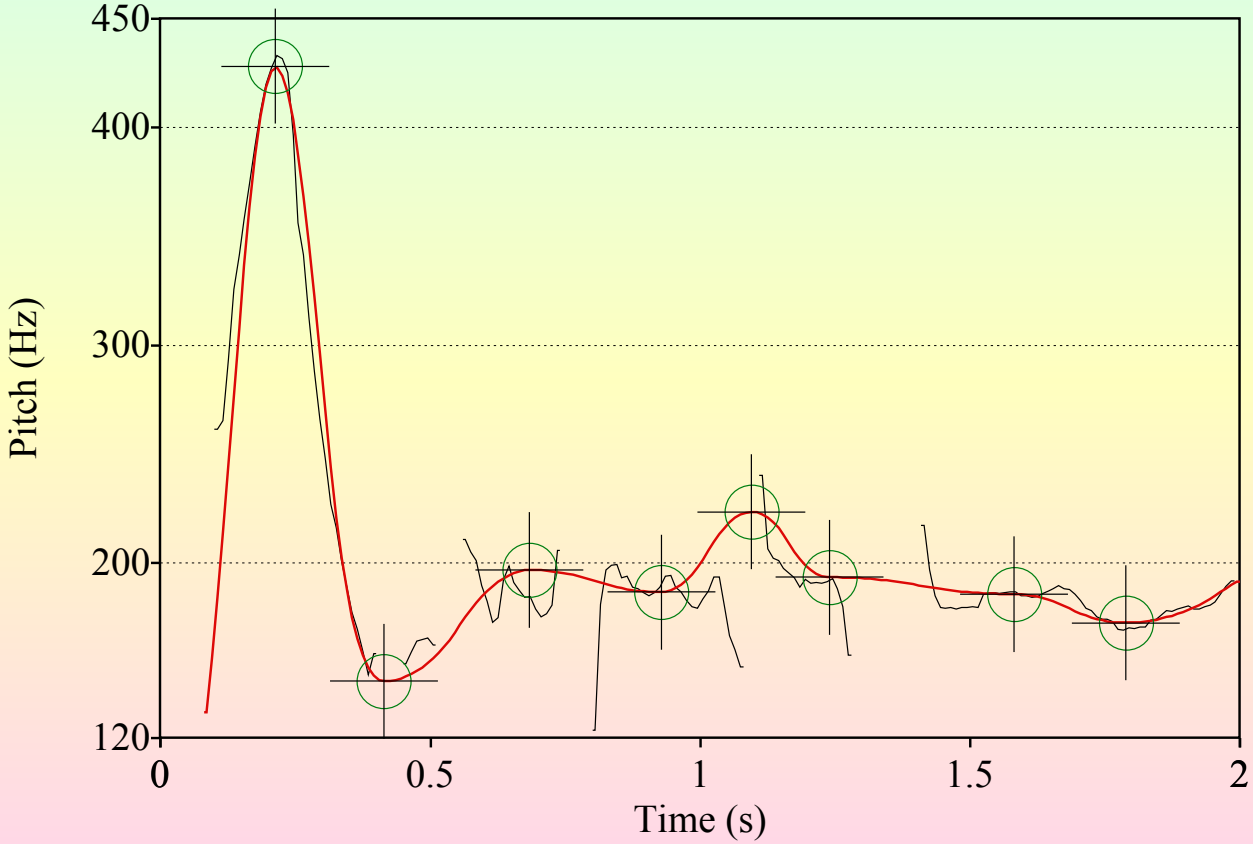

Figure: Momel

Surface Phonological Representation

- INTSINT designed as tool for linguists
  for the symbolic coding of intonation patterns.
  (Hirst & Di Cristo (eds) 1998)

Surface Phonological Representation

- INTSINT designed as tool for linguists
  for the symbolic coding of intonation patterns.
  (Hirst & Di Cristo (eds) 1998)
- Momel and INTSINT are both now implemented as plugin for Praat

INTSINT to Momel

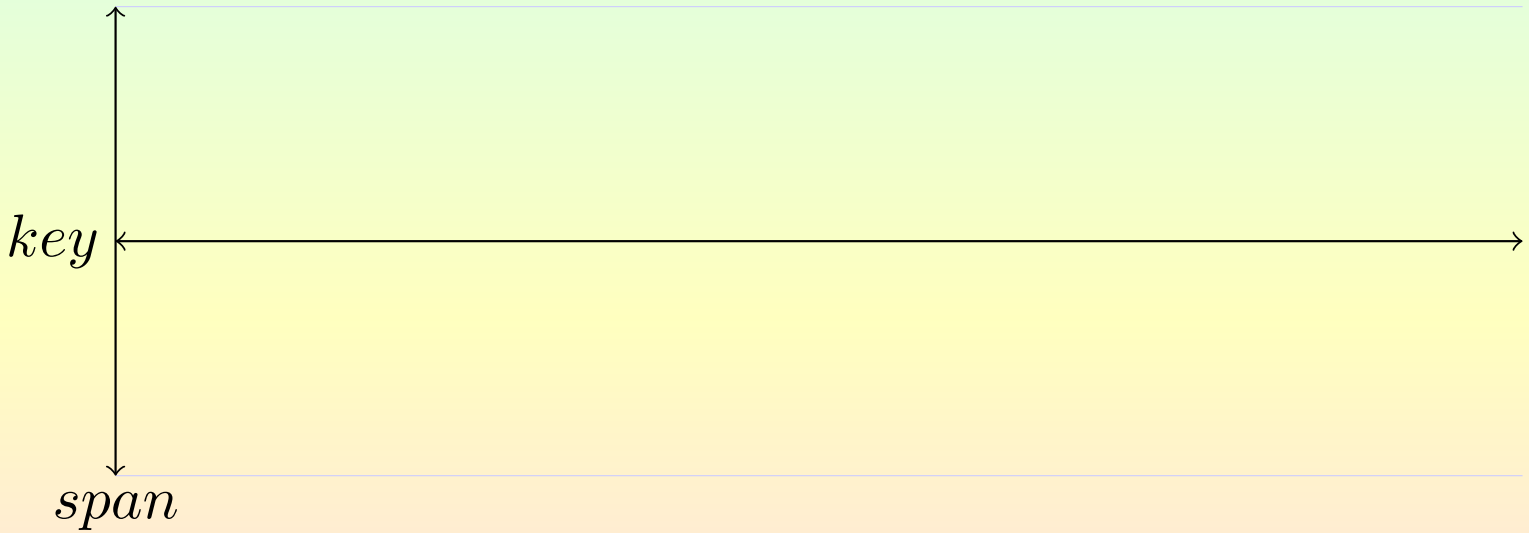

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

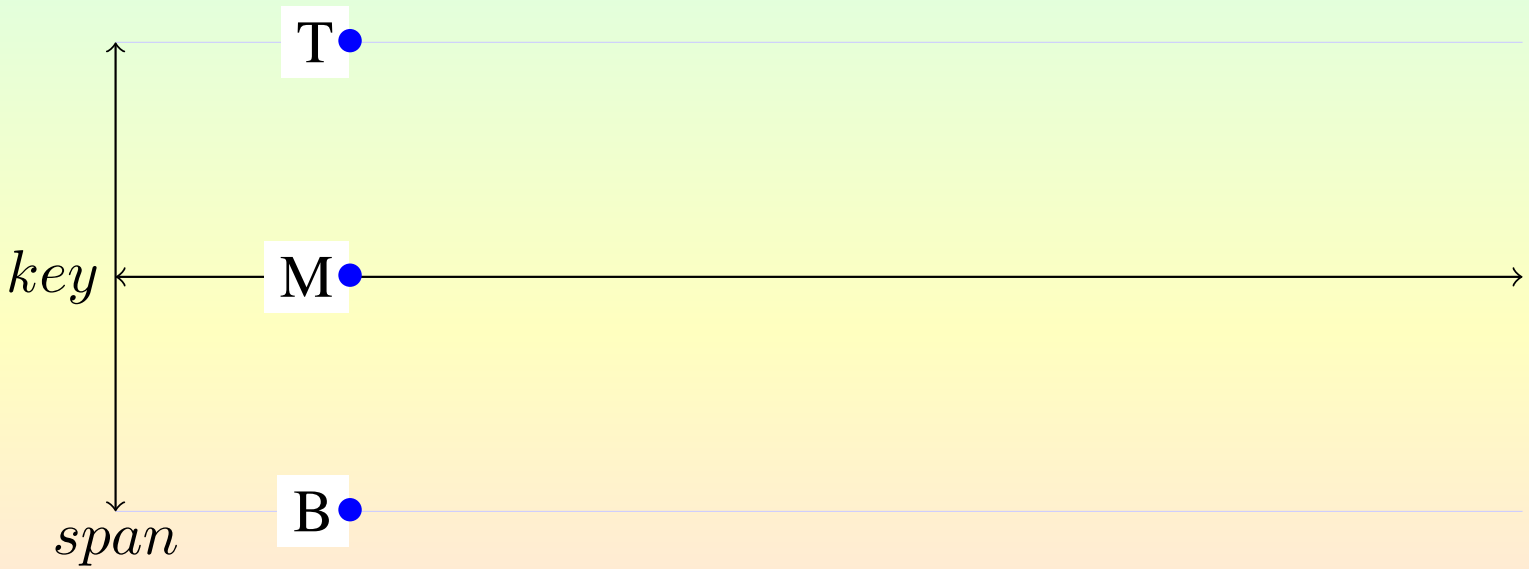

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

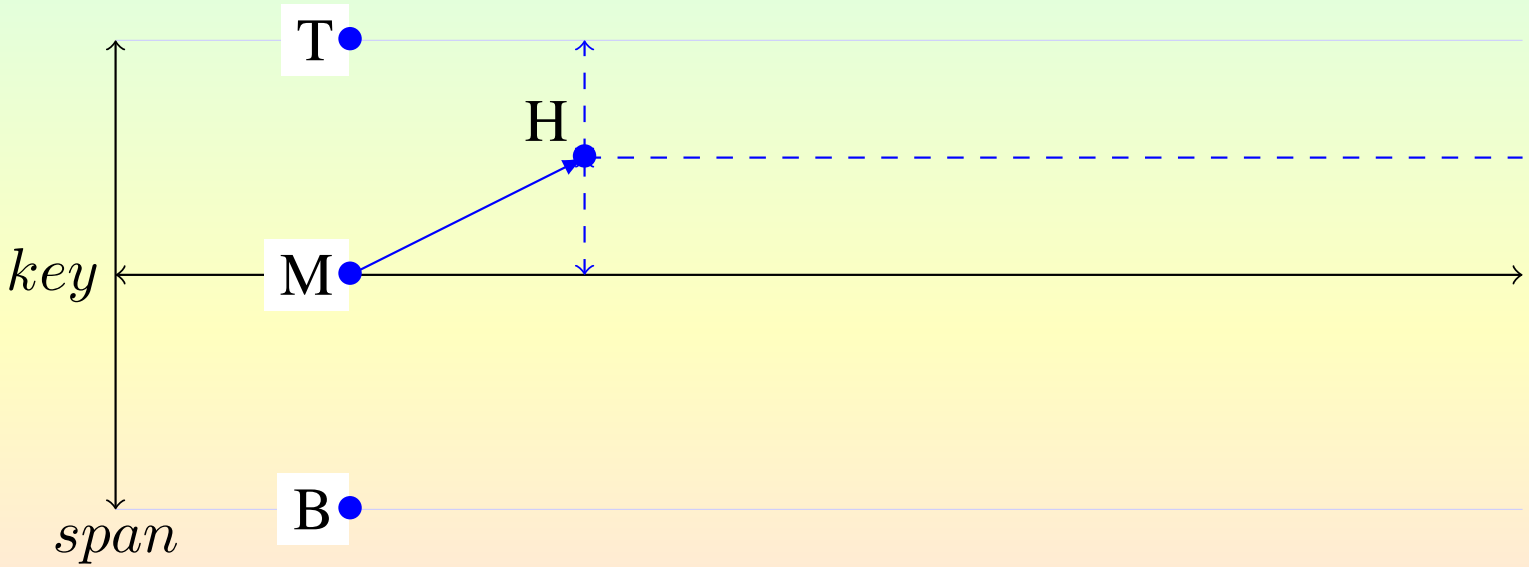

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

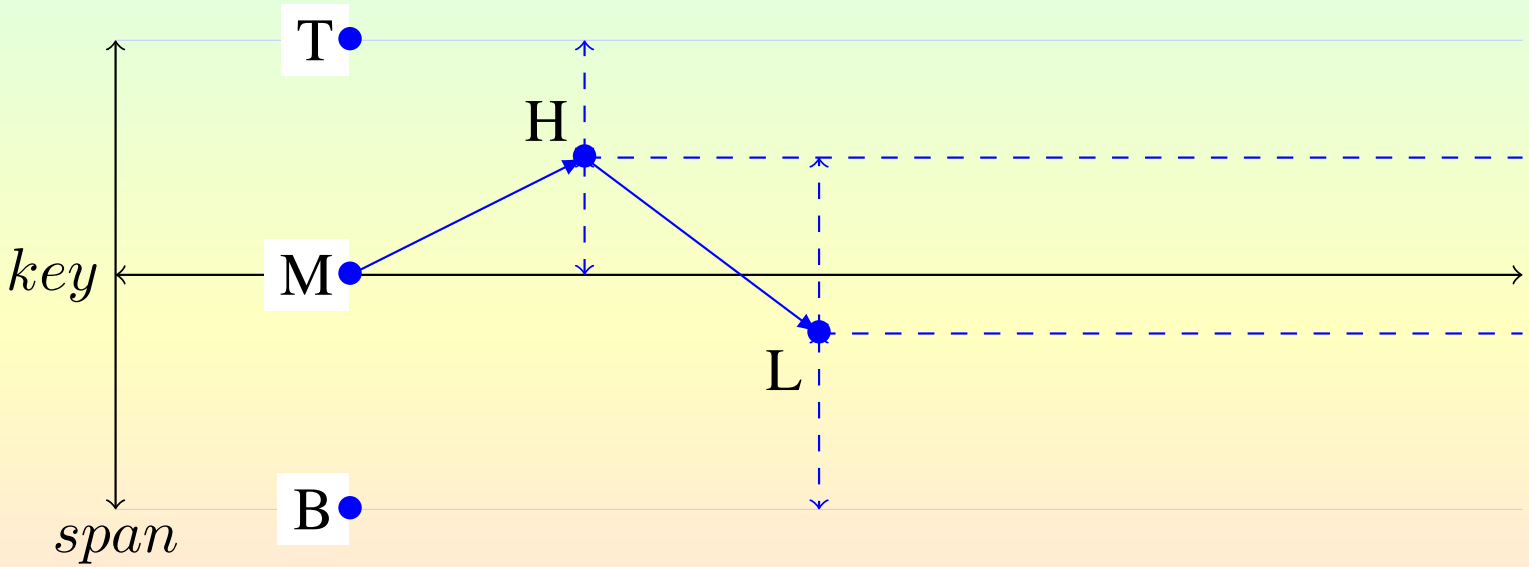

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

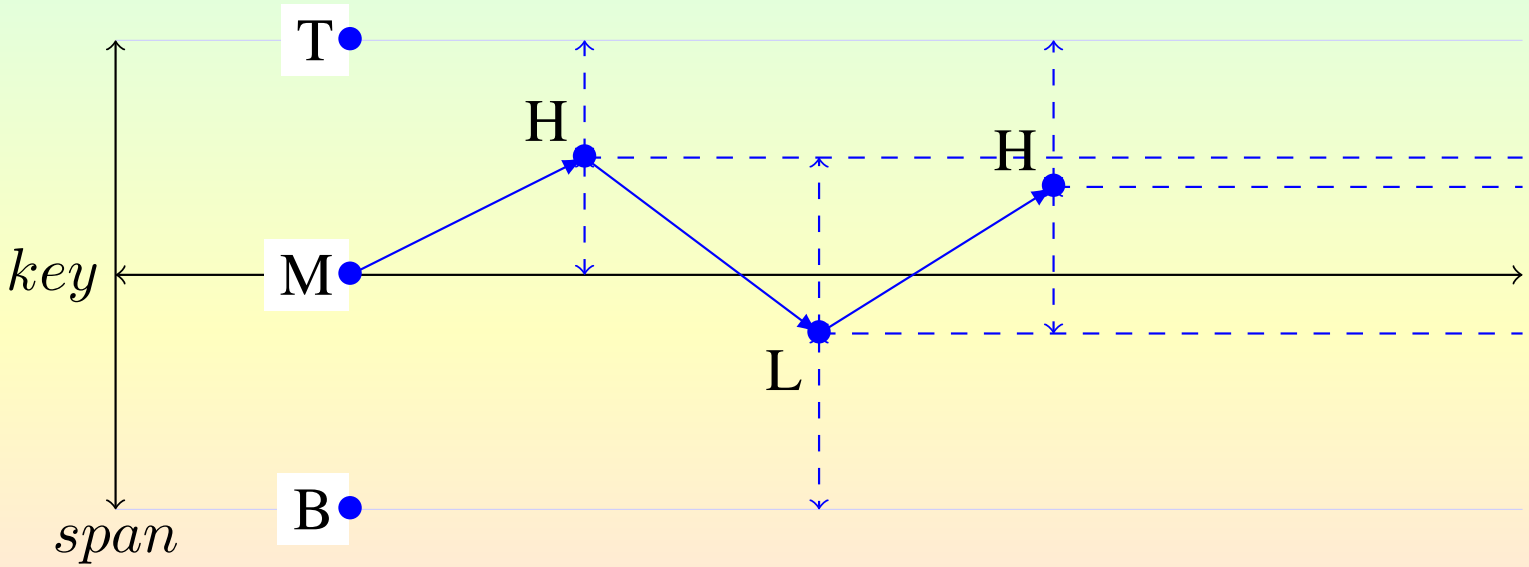

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

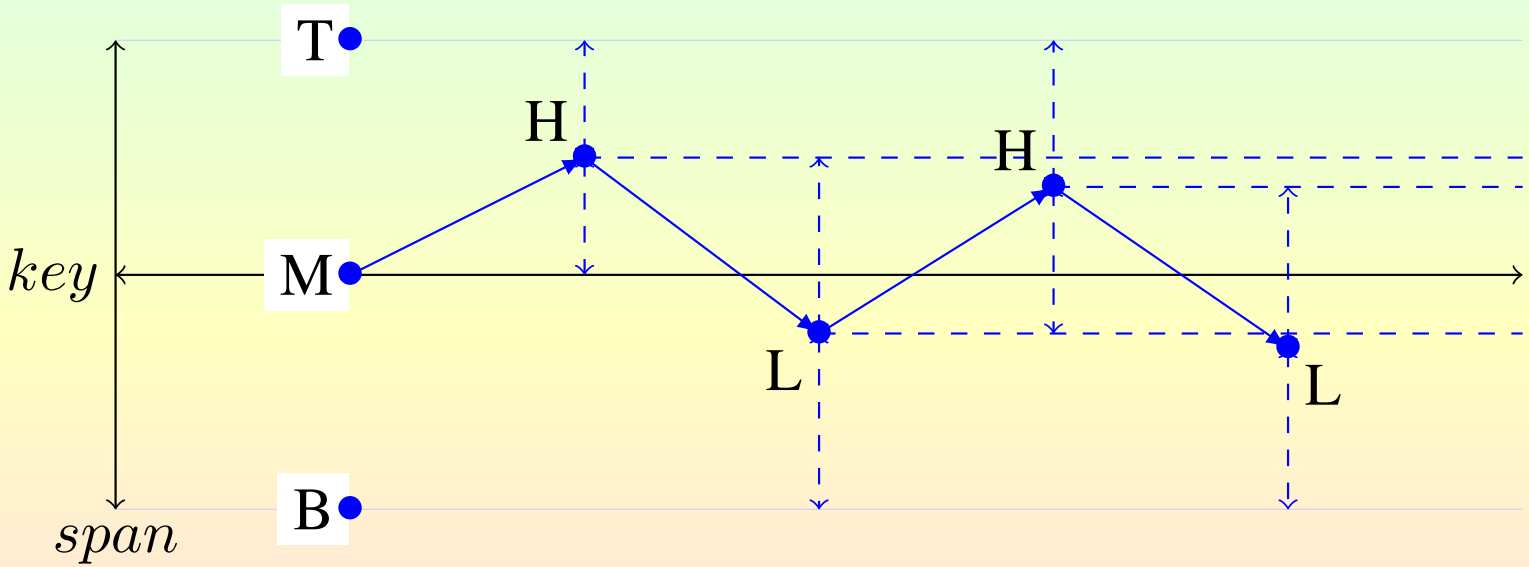

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

INTSINT to Momel

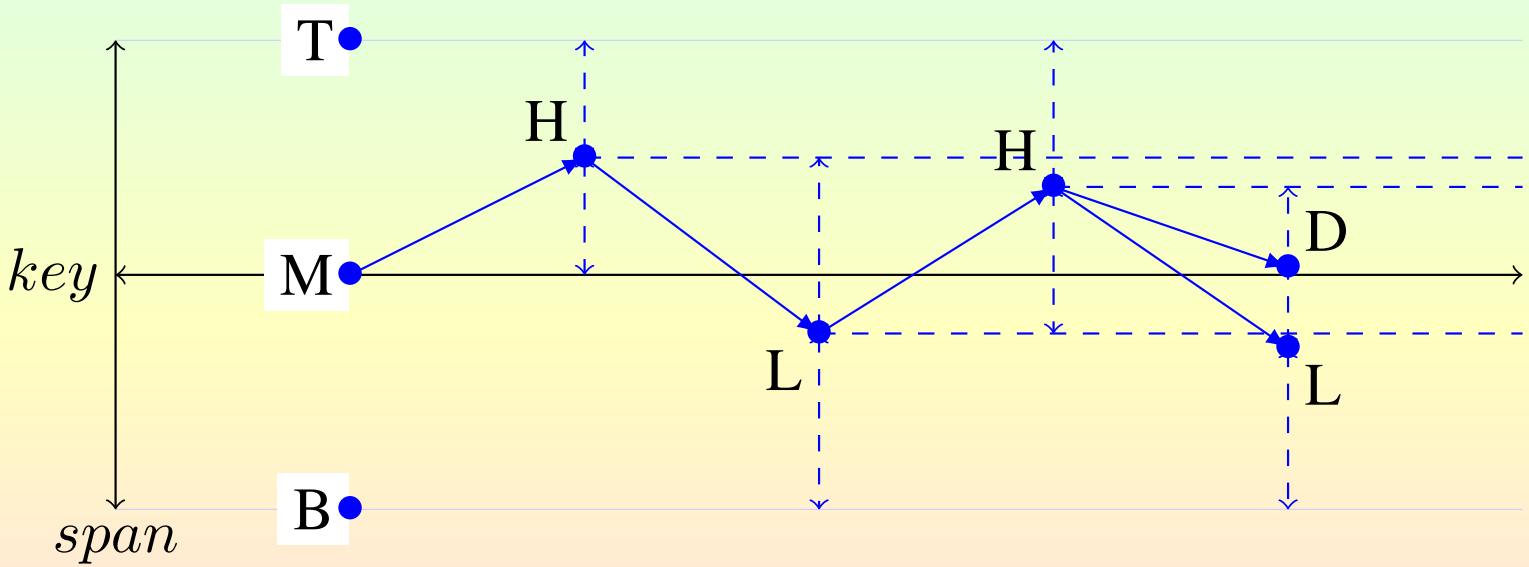

Figure: INTSINT to MoMel defined by 2 parameters key and $span$

Prosodic function

- IF annotation (Hirst 1977, 2005)
- 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
- 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
- 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

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
- analysis by synthesis
- Hirst, D.J. 2011. The analysis by synthesis of speech melody: from data to models. Journal of Speech Sciences 1 (1), 55-83. http://http://www.journalofspeechsciences.org

Analysis by synthesis

Analysis by synthesis

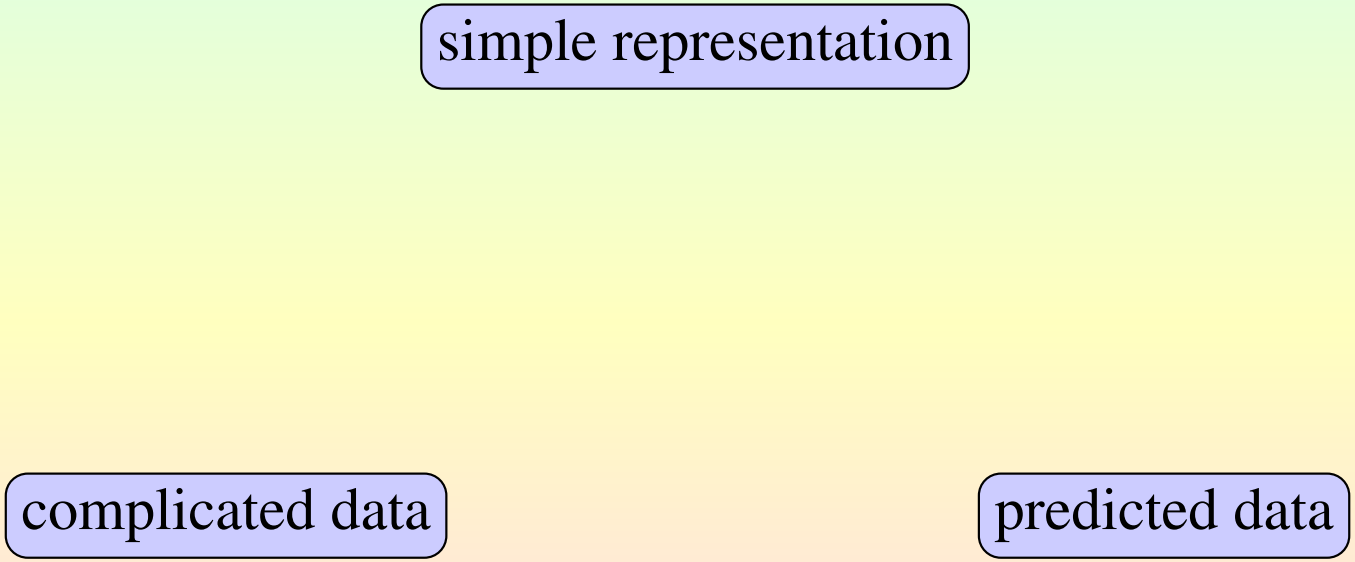

Figure: The Analysis by Synthesis paradigm

Analysis by synthesis

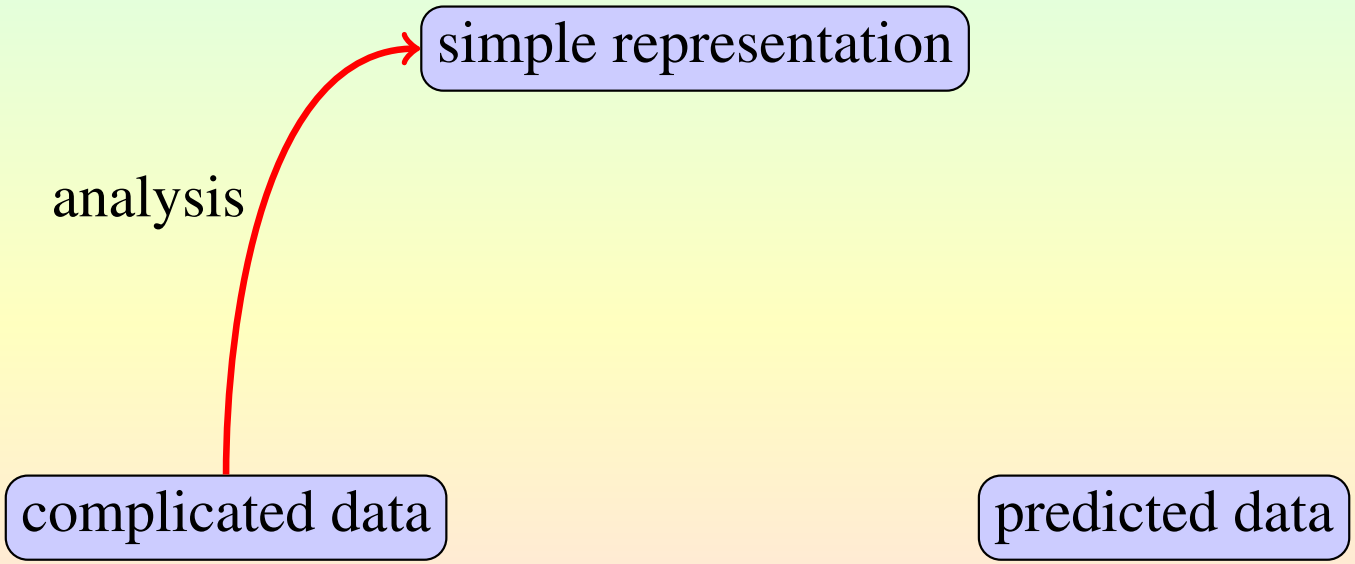

Figure: The Analysis by Synthesis paradigm

Analysis by synthesis

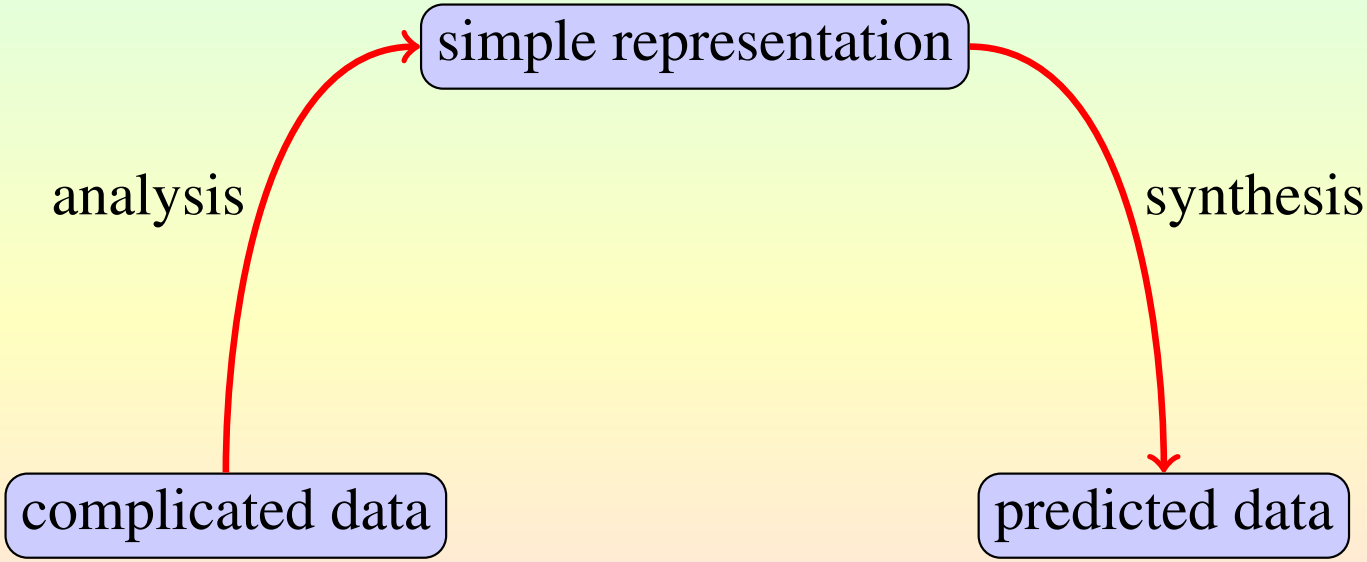

Figure: The Analysis by Synthesis paradigm

Analysis by synthesis

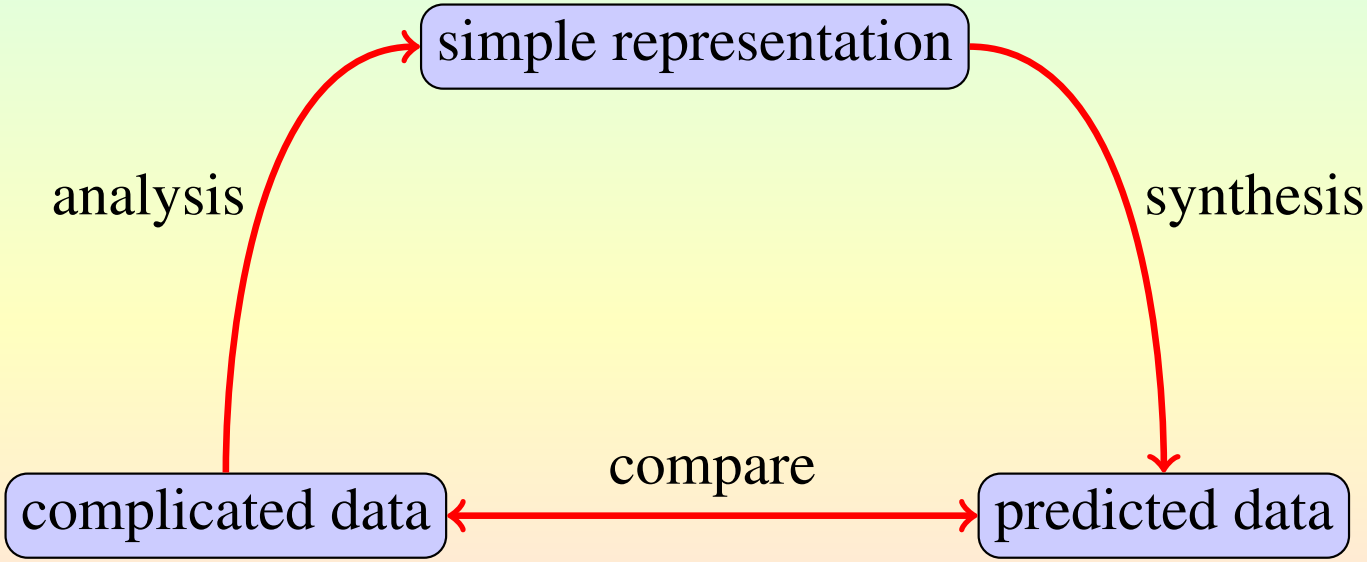

Figure: The Analysis by Synthesis paradigm

What is science?

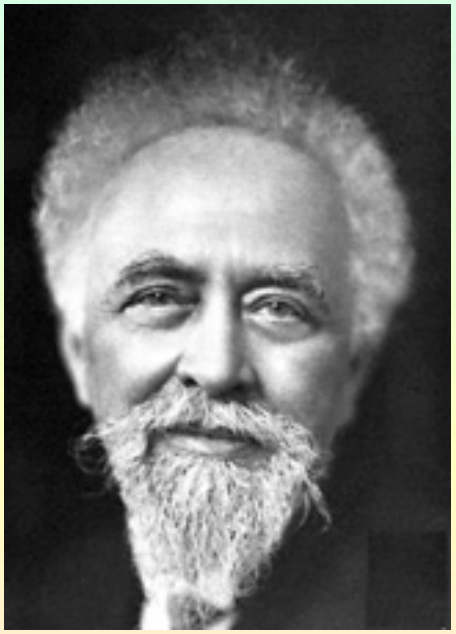

Figure: Jean Baptiste Perrin (1870-1942).

What is science?

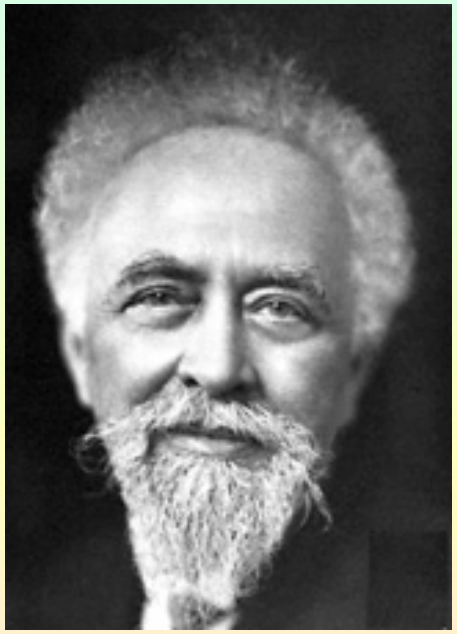

Figure: Jean Baptiste Perrin (1870-1942).

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