MultilingualWeb, Luxembourg, 16th of March, 2012: Results of a theme session on

Semantic Resources and Machine Learning for Quality, Efficiency and Personalisation of Accessing Relevant Information over Language Borders

(different languages and
different uses of a same language)

Participants

- Timo Honkela, Aalto University (rapporteur)
- Peter Schmitz, Publications Office of the EU
- Elena Montanes, Oviedo University
- Tasos Koutoumanos, AgroKnow Tech., Greece
- Corinne Frappart, Publications Office of the EU
- Poul Andersen, WEB translation unit, EU Commission
- Ghassan Haddad, Facebook
- Spyridon Pilos, Language applications, European Commission
- Jose Emilio Labra Gayo, University of Oviedo, Spain
- Maria Pia Montoro, Intrasoft International, Luxembourg
- Daaniel Garcia Magarinos, European Central Bank

Quality and consistency versus accessibility and contextual appropriateness of terminology

- Terms good for experts in different domains versus laypersons
- Case: “member state” versus “EU country”
- Case: “human trafficking” versus “modern slavery”
- Case: Bank note security features
  - A thesaurus was created as a mapping from technical terms to colloquial language
    (“iridescent stripe” to “glossy stripe”)
- Case: legislation (Asturias region in Spain): mapping of colloquial terms to official terms, new project: library of congress in Chile

Quality and consistency versus accessibility and contextual appropriateness of terminology

- Convergent and divergent processes in language use
  - Ontologies: carefully crafted resources that require considerable resources for implementation and use
  - Folksonomies: resources that provide information on the variation and are constructed by the crowds

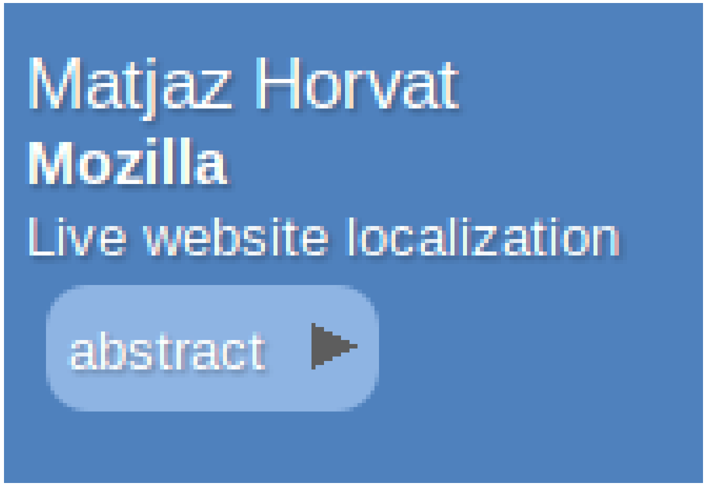
    > Possibility to model the crowdsourced data using machine learning techniques

Multilingual contents and thesauri: trust and quality

- Use of EU-generated resources such as
  - Eurovoc
  - JRC-Names
- Importance of linked open data (LOD)
  - Choosing keywords from a controlled vocabulary
  - Connecting different term versions with an ontology (or folksonomy)
  - Determining a proper contexts using LOD
- Multilingual content: provenance of data
- Quality assurance of LOD

Effect of context in translation: need for context-rich representations

- Often the variation in translation of terminology stems from contextual factors
- It would be important to store enough contextual information in order to facilitate appropriate choices



Matjaz Horvat

Mozilla

Live website localization

Social and cognitive levels of language use

- Push and pull of terminology
  - Regulation and market economy of language
- Different levels of expertise
  - Experts in different domains versus laypersons
- Take home messages:
  - Variation among language in
    conceptual structures
    (challenges for ontology translation)
  - Semantic variation among languge users

Space under Construction

Language-Specific Spatial Categorization

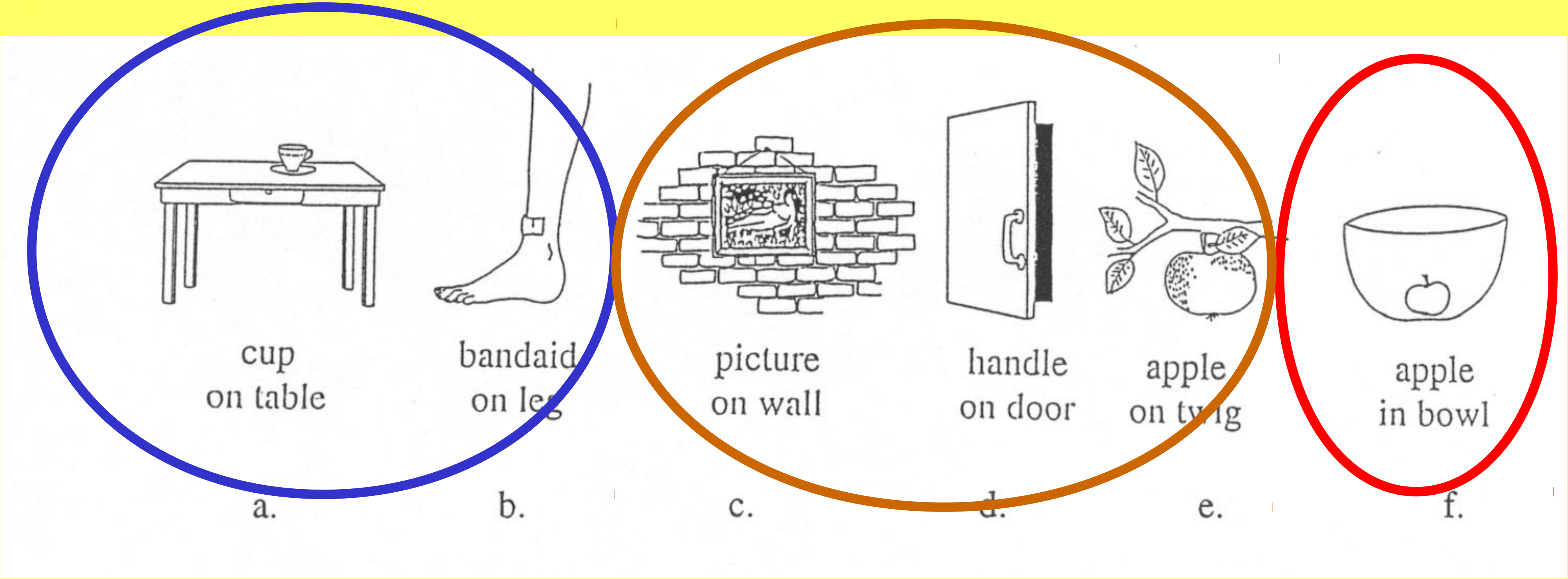
In First Language Acquisition

Melissa Bowerman

Max Planck Institute for Psycholinguistics

Lund University Cognitive Science
2003

DUTCH

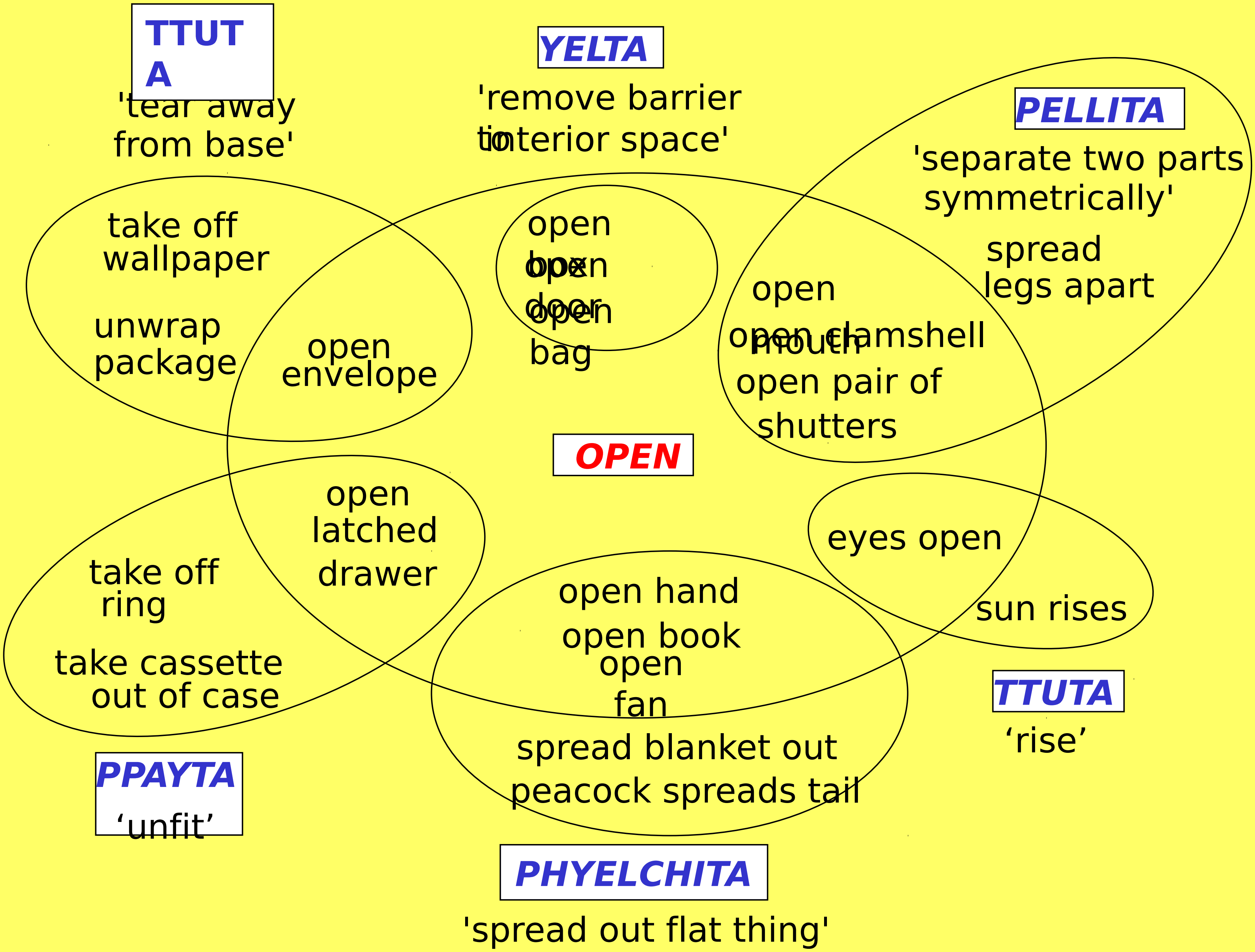


OP

AAN

IN

Categorization of `opening' in **English** and **Korean**.



TTUT A
'tear away from base'

YELTA
'remove barrier to interior space'

PELLITA
'separate two parts symmetrically'

take off wallpaper

unwrap package

open envelope

open box

open door

open bag

open mouth

open clamshell

open pair of shutters

spread legs apart

OPEN

open latched drawer

take off ring

take cassette out of case

eyes open

sun rises

open hand

open book

open fan

spread blanket out

peacock spreads tail

TTUTA
'rise'

PPAYTA
'unfit'

PHYELCHITA
'spread out flat thing'

(Pye 1995, 1996)

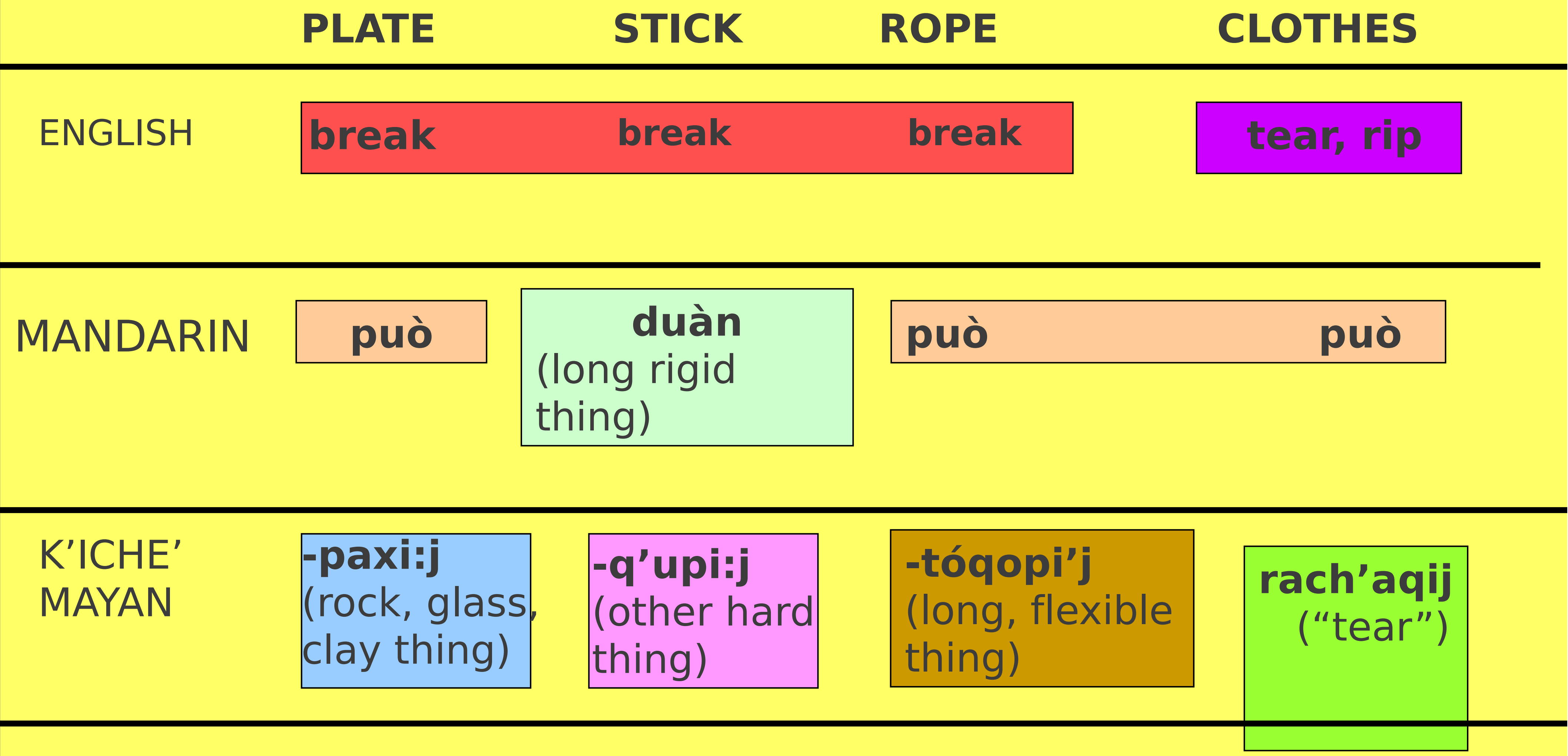
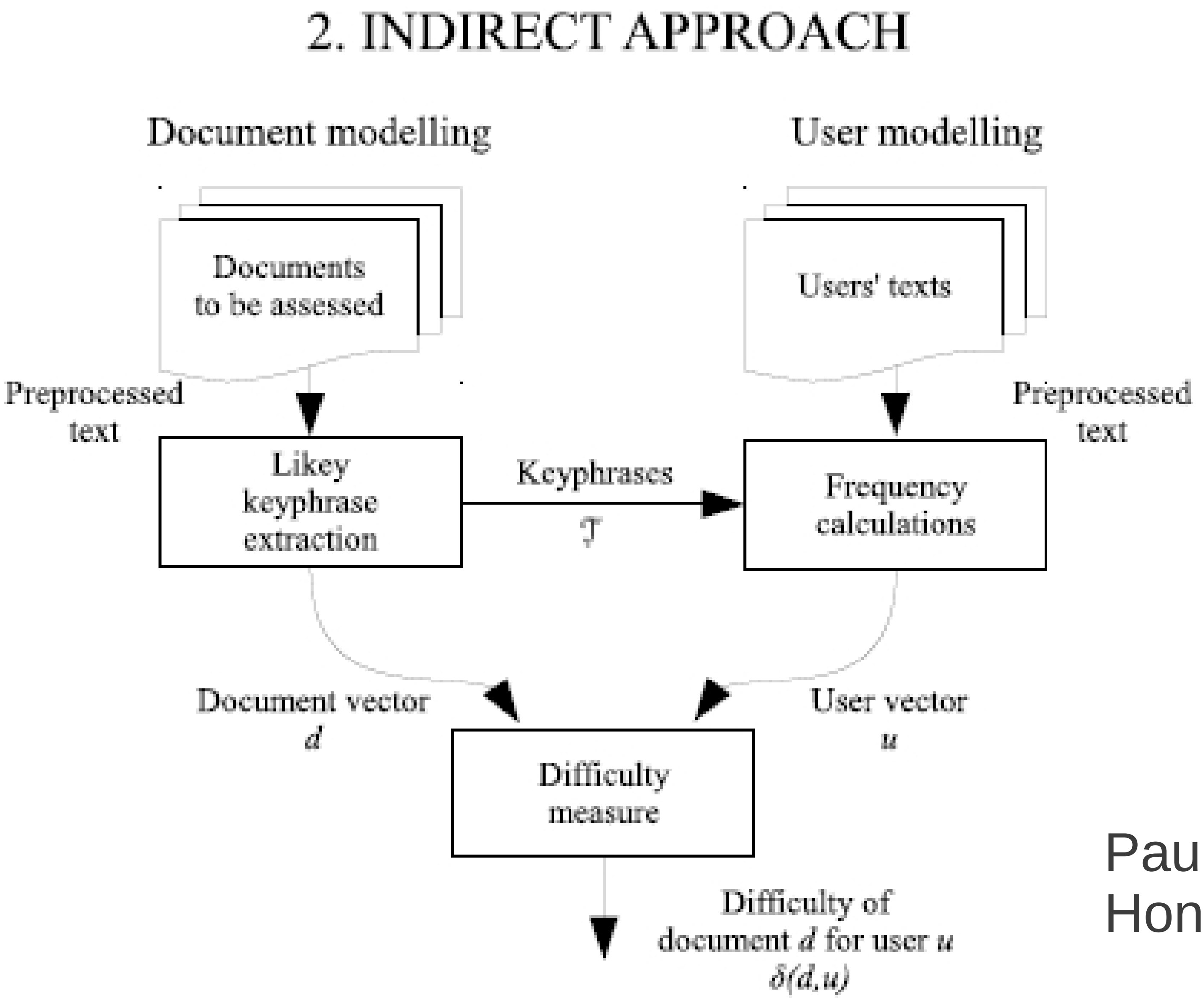


	PLATE	STICK	ROPE	CLOTHES
ENGLISH	**break**	**break**	**break**	**tear, rip**
MANDARIN	**può**	**duàn** (long rigid thing)	**può**	**può**
K'ICHE' MAYAN	**-paxi:j** (rock, glass, clay thing)	**-q'upi:j** (other hard thing)	**-tóqopi'j** (long, flexible thing)	**rach'aqij** ("tear")

http://www.mpi.nl/people/bowerman-melissa

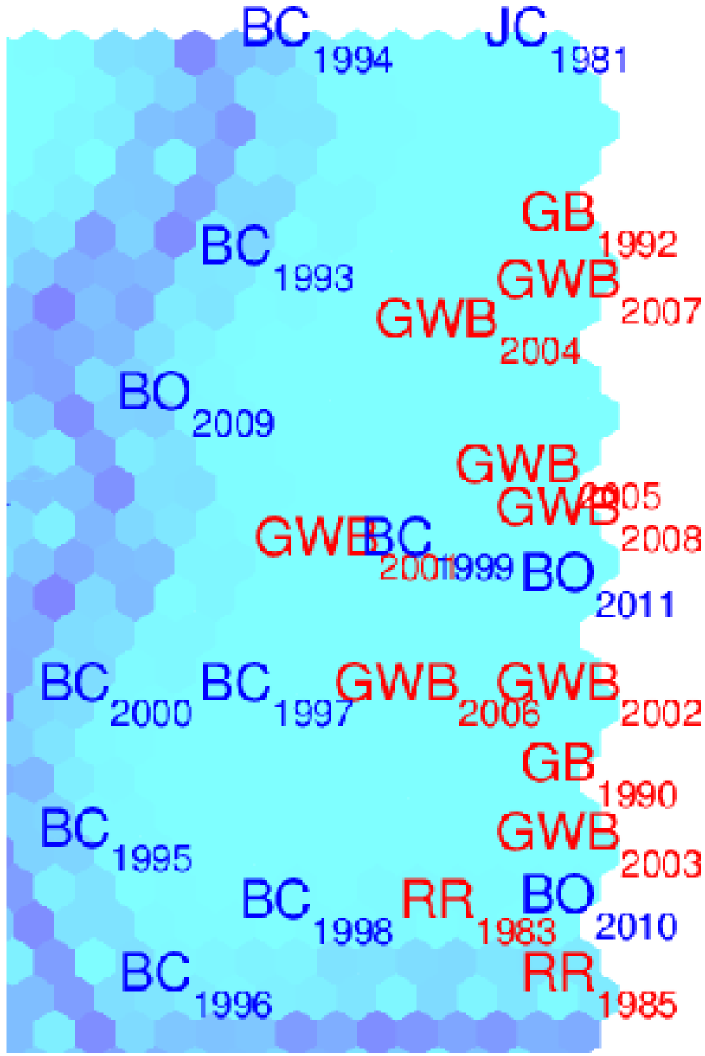
http://www.mpi.nl/people/bowerman-melissa/publications

User-specific difficulty measure



Paukkeri, Ollikainen & Honkela, submitted

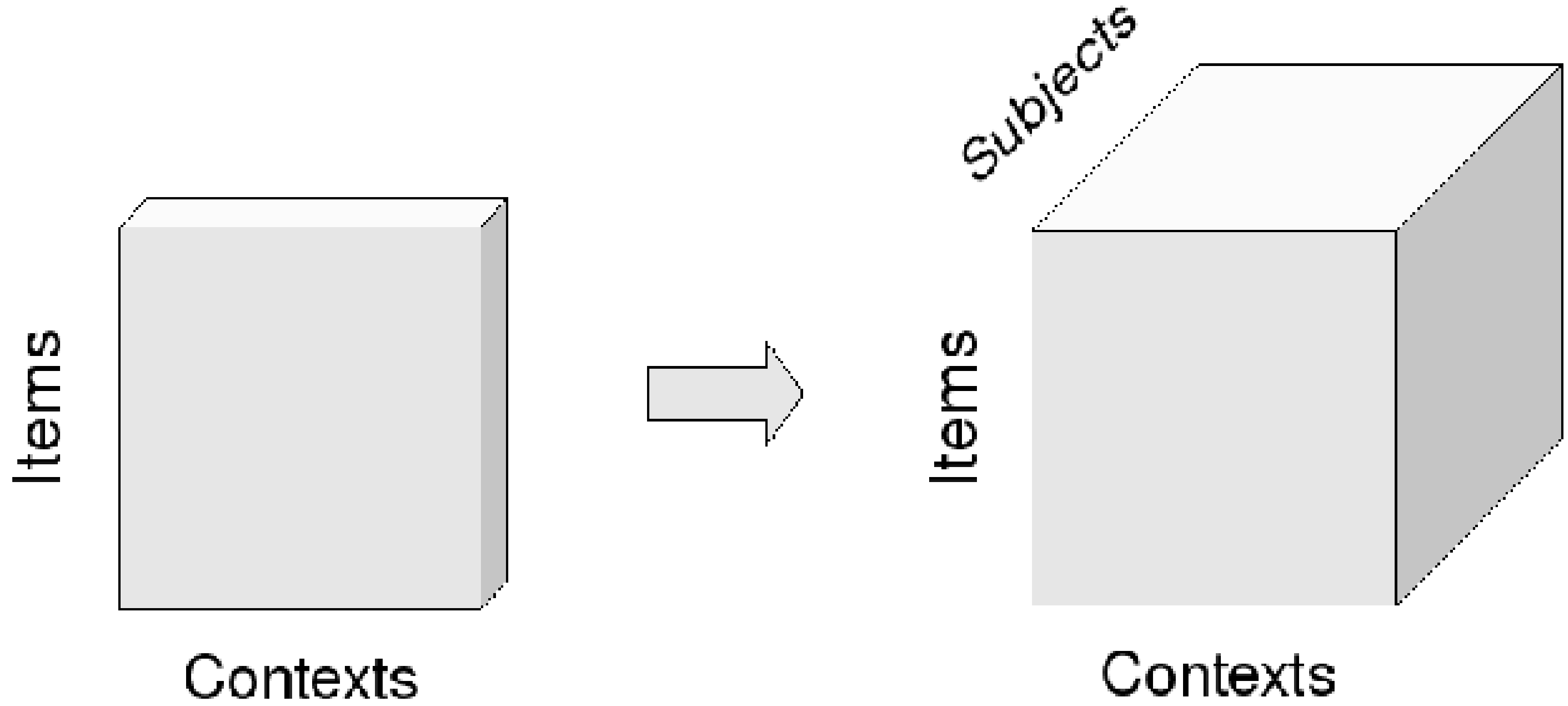
GICA analysis: Word 'health' in State of the Union Addresses



GICA: Grounded Intersubjectivity Concept Analysis

Timo Honkela, Juha Raitio, Krista Lagus, Ilari T. Nieminen, Nina Honkela, and Mika Pantzar. **Subjects, objects and contexts: Using GICA method to quantify epistemological subjectivity**. In *Proceedings of IJCNN 2012, International Join Conference on Neural Networks,* to appear.

Core of GICA: Subject-Object-Context Tensors



Timo Honkela, Nina Janasik, Krista Lagus, Tiina Lindh-Knuutila, Mika Pantzar, and Juha Raitio. GICA: Grounded intersubjective concept analysis - a method for enhancing mutual understanding and participation. Technical Report TKK-ICS-R41, AALTO-ICS, ESPOO, December 2010.

http://users.ics.tkk.fi/tho/info/TKK-ICS-R41.shtml http://users.ics.tkk.fi/tho/publications.shtml

Guidelines are needed on how to publish data in multiple languages

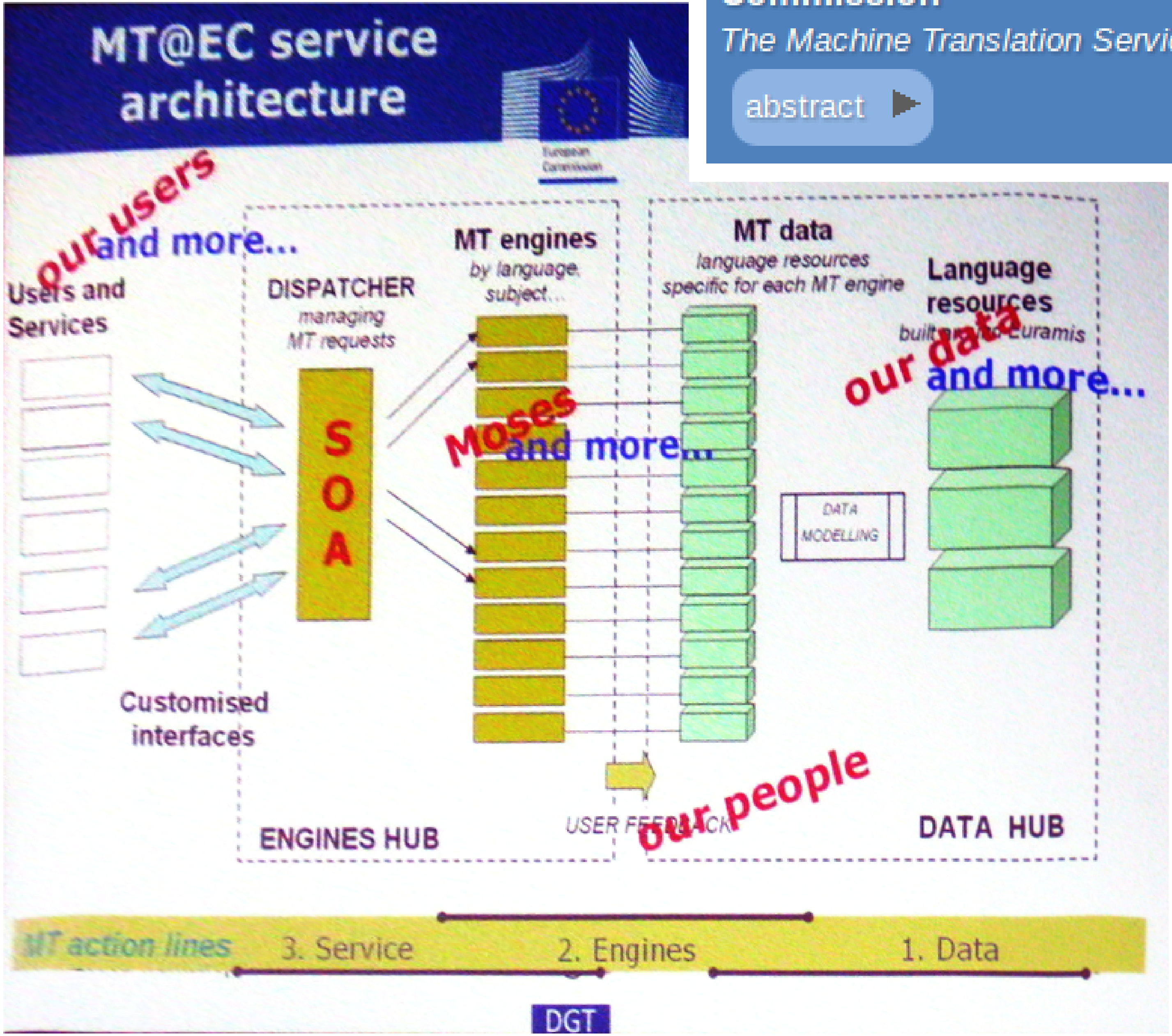
- Different versions in different languages
  - Alternative language versions
  - A standard way of describing how how different versions are related to each other
- Case FAO: Translations should refer back to the original documents

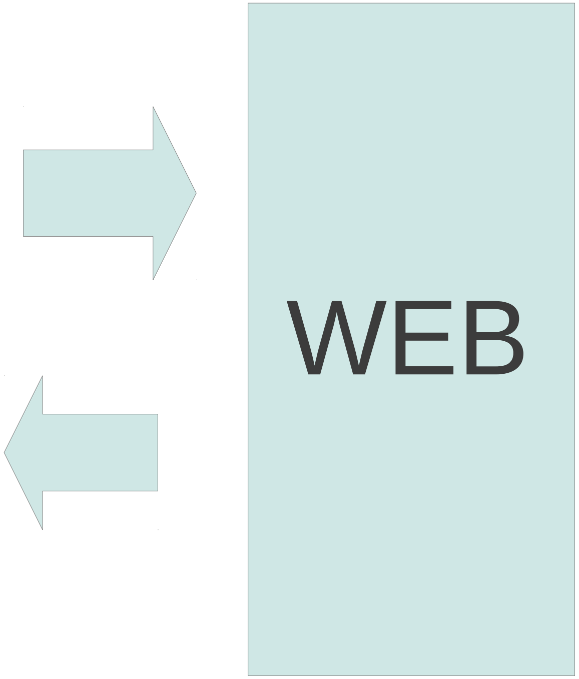
Spyridon Pilos

Directorate General for Translation - European Commission

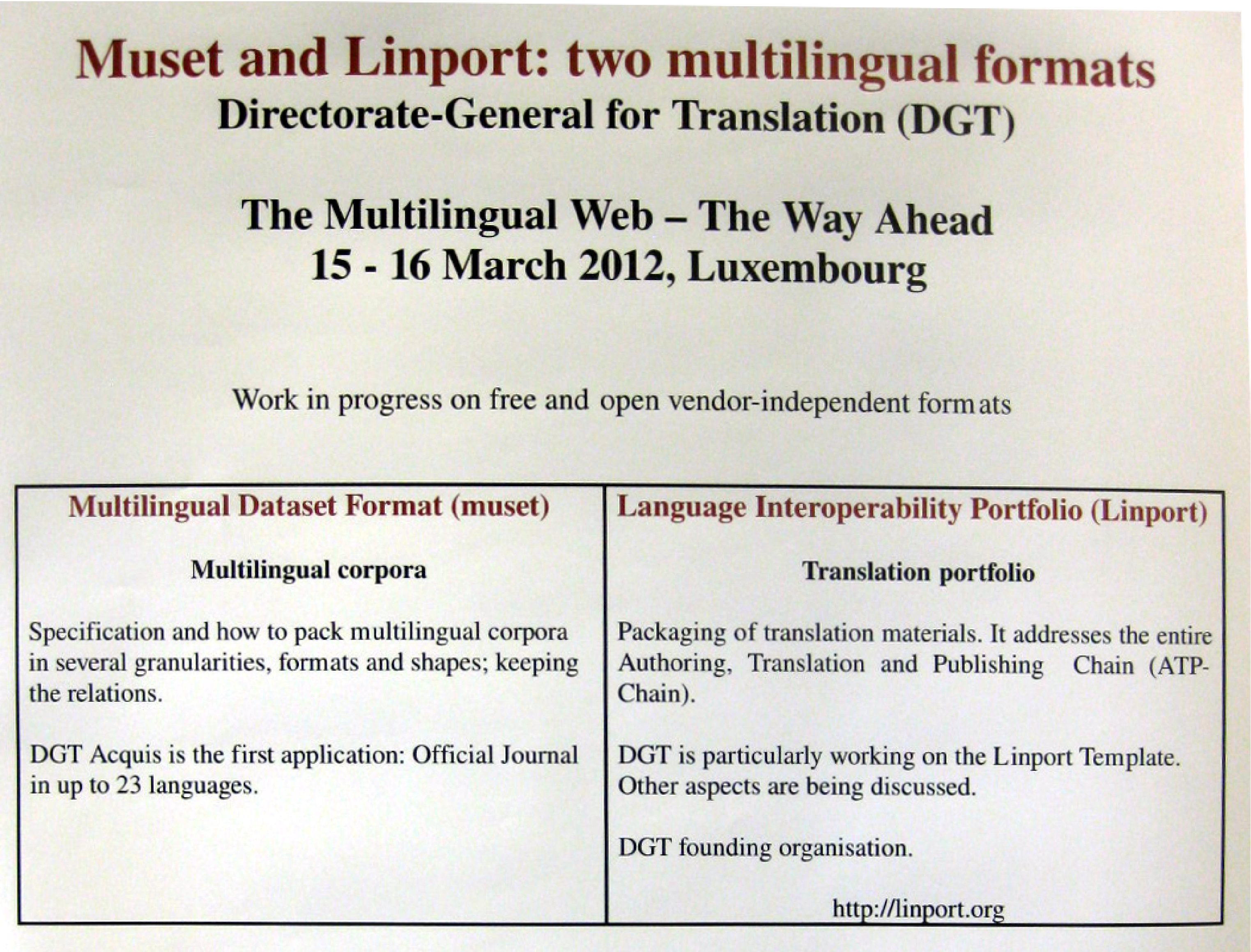
The Machine Translation Service of the European Commission

abstract ▶





Linport has related objectives



Muset and Linport: two multilingual formats

Directorate-General for Translation (DGT)

The Multilingual Web – The Way Ahead
15 - 16 March 2012, Luxembourg

Work in progress on free and open vendor-independent formats

Multilingual Dataset Format (muset)	Language Interoperability Portfolio (Linport)
Multilingual corpora	**Translation portfolio**
Specification and how to pack multilingual corpora in several granularities, formats and shapes; keeping the relations. DGT Acquis is the first application: Official Journal in up to 23 languages.	Packaging of translation materials. It addresses the entire Authoring, Translation and Publishing Chain (ATP-Chain). DGT is particularly working on the Linport Template. Other aspects are being discussed. DGT founding organisation. http://linport.org