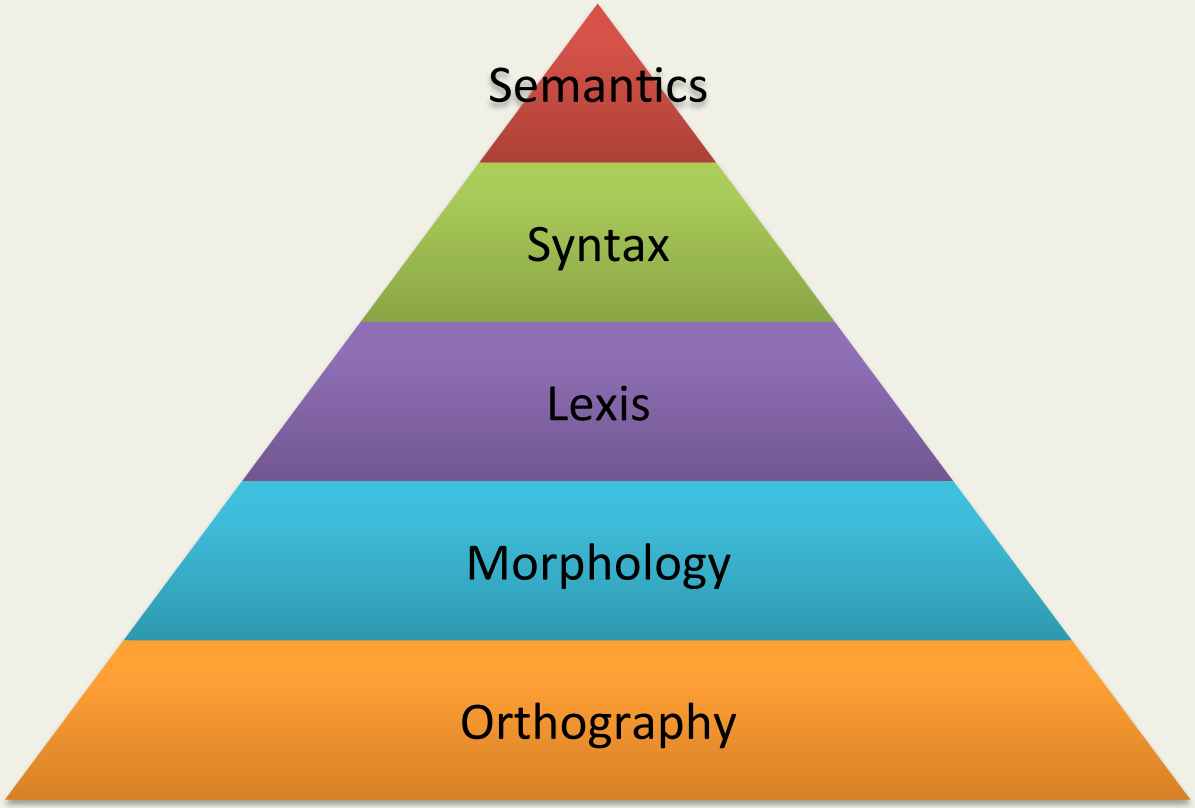
Approaches to Machine Translation: Rule-based, Statistical and Hybrid

SMT CHALLENGES

SMT challenges
Language linguistics



Orthographic Challenges

Spelling mistakes and typographical errors

Even minor, they convert an existent word in the training corpus into an out-of-vocabulary word.

Truecasing and capitalization

It is common in SMT to lowercase all training and testing data in order to avoid orthographic mismatchings.

Normalization

Some words can be usually written in different ways, leading to orthographic differences with respect to the trained corpus.

(De-)Tokenization

Splitting a stream of text up into appropriate tokens to facilitate the input for further processing a text.

Transliteration

Conversion of text strings from one orthography to another, while preserving the phonetics in both languages.

Orthographic Challenges

Spelling mistakes and typographical errors

e.g. *conecting** instead of *connecting*

Truecasing and capitalization

e.g. From *obama* to *Obama*

Normalization

e.g. *Dr. dr. Dr dr PhD*

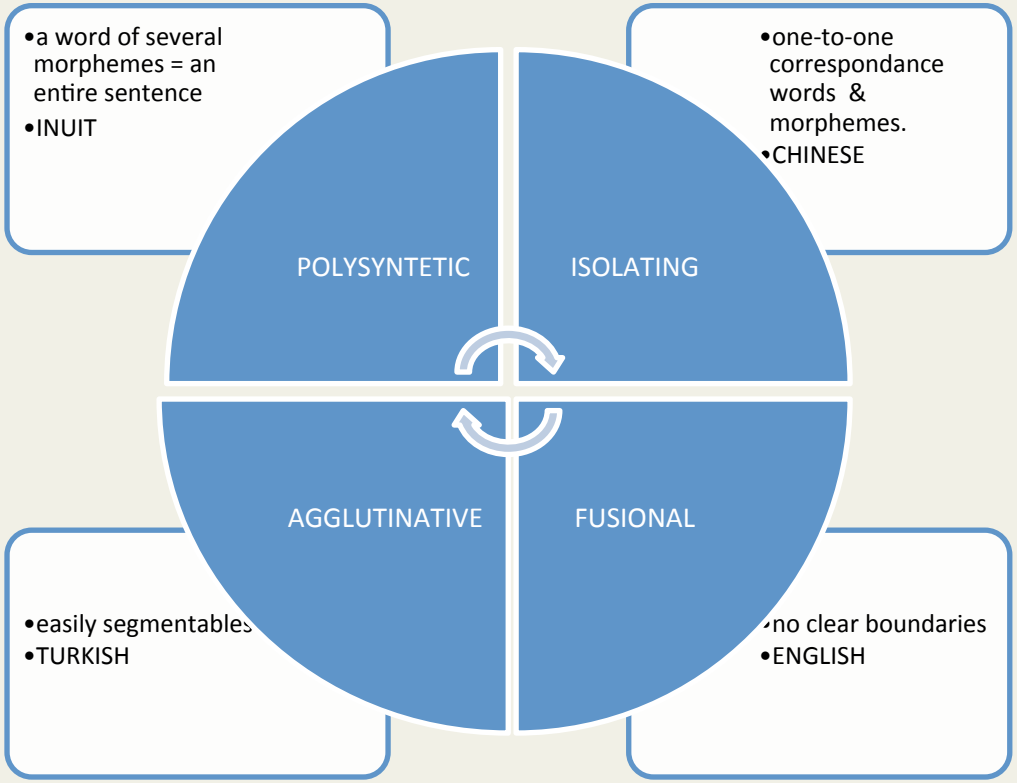
(De-)Tokenization

e.g. *“ Hello “* vs *“Hello”*

Transliteration

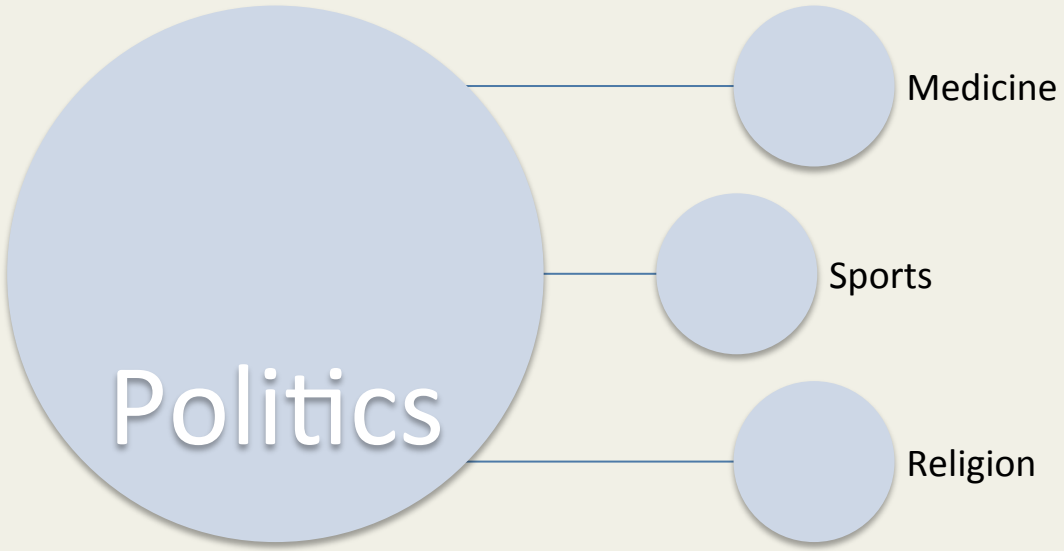
e.g. *"Ελληνική Δημοκρατία" 'Hellenic Republic'* can be transliterated as *"Ellēnikē Dēmokratia"*

Morphological Challenges



Lexical Challenges

- Unknown words
  - Words that do not appear in the training set
  - In-domain vs Out-domain



Syntactic Challenges

SYNTAX

principles and rules for constructing sentences in natural language

Not included in PBSMT ➡ phrases are just sequences of words with no structure.

Consequence: **word reordering errors** (specially when translating into more fixed-order languages like English.

Semantic challenges

- Often we start with ***lexical semantics***, the meanings of words
  - » To translate a word correctly, we need to know what it means
- Language generally follows the ***principle of compositionality***
  - » Meaning of a complex expression is a function of its parts
  - » To translate a sentence correctly, we need to understand the objects and their relationships

Question

(Mark all possible answers)

- Syntactic challenges include long reorderings:
  - When source and target languages have different structures Subject-Verb-Object and Subject-Object-Verb
  - When adjectives and nouns follow different orders
  - When languages involved in the translation are derived from the same family
  - When word ordering in one or both languages are flexible

Question

(Mark all possible answers)

- Syntactic challenges include long reorderings:
  - **When source and target languages have different structures Subject-Verb-Object and Subject-Object-Verb**
  - When adjectives and nouns follow different orders
  - When languages involved in the translation are derived from the same family
  - **When word ordering in one or both languages are flexible**

Linguistic level	Challenge	Main related works
Orthography	Spelling	[Bertoldi et al. 2010][Farrús et al. 2011]
	Truecasing/Capitalization	[Lita et al. 2003][Wang et al. 2006]
	Normalization	[Riesa et al. 2006][Aw et al. 2006] [Diab et al. 2007][Kobus et al. 2008]
	Tokenization	[Farrús et al. 2011][El Kholy and Habash 2012]
	Transliteration	[Boas 2002][Virga and Khudanpur 2003] [Kondrak et al. 2003][Zhang et al. 2004] [Kondrak 2005][Mulloni and Pekar 2006] [Kumaran and Kellner 2007][Mitkov et al. 2007] [Istvan and Shoichi 2009][Nakov and Ng 2009]
Morphology	Inflections	[Brants 2000][Ueffing and Ney 2003] [Creutz and Lagus 2005][Minkov et al. 2007] [Koehn and Hoang 2007][Virpioja et al. 2007] [Avramidis and Koehn 2008][de Gispert et al. 2009] [El-Kahlout and Oflazer 2010] [Bojar and Tamchyna 2011][Green and DeNero 2012] [Formiga et al. 2012][Rosa et al. 2012]
Lexis	Unknown words	[Knight and Graehl 1998] [Al-Onaizan and Knight 2002] [Koehn and Knight 2003][Fung and Cheung 2004] [Shao and Ng 2004][Langlais and Patry 2007] [Mirkin et al. 2009][Marton et al. 2009][Li et al. 2010] [Huang et al. 2011][Zhang et al. 2012]
	Spurious words	[Fraser and Marcu 2007][Li and Yarowsky 2008] [Menezes and Quirk 2008]
Syntax	Word reordering	[Wu 1997][Alshawi et al. 2000] [Menezes and Richardson 2001] [Yamada and Knight 2002][Aue et al. 2004] [Galley et al. 2004][Ringger et al. 2004] [Xia and McCord 2004][Chiang 2005] [Collins et al. 2005][Ding and Palmer 2005] [Quirk et al. 2005][Simard et al. 2005] [Zhang and Gildea 2005][Galley et al. 2006] [Liu et al. 2006][Huang et al. 2006] [Langlais and Gotti 2006][Smith and Eisner 2006] [Turian et al. 2006][Birch et al. 2007][Li et al. 2007] [Zhang et al. 2007][Wang et al. 2007][Cowan 2008] [Elming 2008][Graehl et al. 2008] [Li and Yarowsky 2008][Badr et al. 2009] [Genzel 2010][Shen et al. 2010] [Khalilov and Fonollosa 2011][Bach 2012] [Germann 2012]
Semantics	Sense disambiguation	[Garcia-Varea et al. 2001][Chiang 2005] [Bangalore et al. 2007][Carpuat and Wu 2007] [Chan et al. 2007][Carpuat and Wu 2008] [Shen et al. 2009][Wu and Fung 2009] [España-Bonet et al. 2010][Haque 2011] [Banchs and Costa-jussà 2011][Banarescu et al. 2013]

Costa-jussà M. R. and Farrús, M.
Statistical Machine Translation Enhancements through Linguistic Levels: A Survey
ACM Computing Surveys, Volume 46 Number 3 pages 42, 2014

Next

Rule-based MT