MACHINE TRANSLATION IN A NUTSHELL
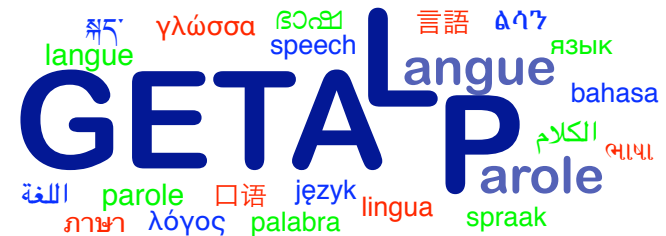
Hervé Blanchon

Laboratoire LIG

Équipe GETALP

herve.blanchon@univ-grenoble-alpes.fr

Outline

- Introduction
- Human Translation
- Machine Translation
- Translation Niches
    - Re-creation
    - Localization
    - Diffusion or Dissemination
    - Acquisition or Assimilation
- How to Describe a MT System
    - Operational Architecture
    - Linguistic Architecture
    - Computational Architecture
- Conclusion
- References
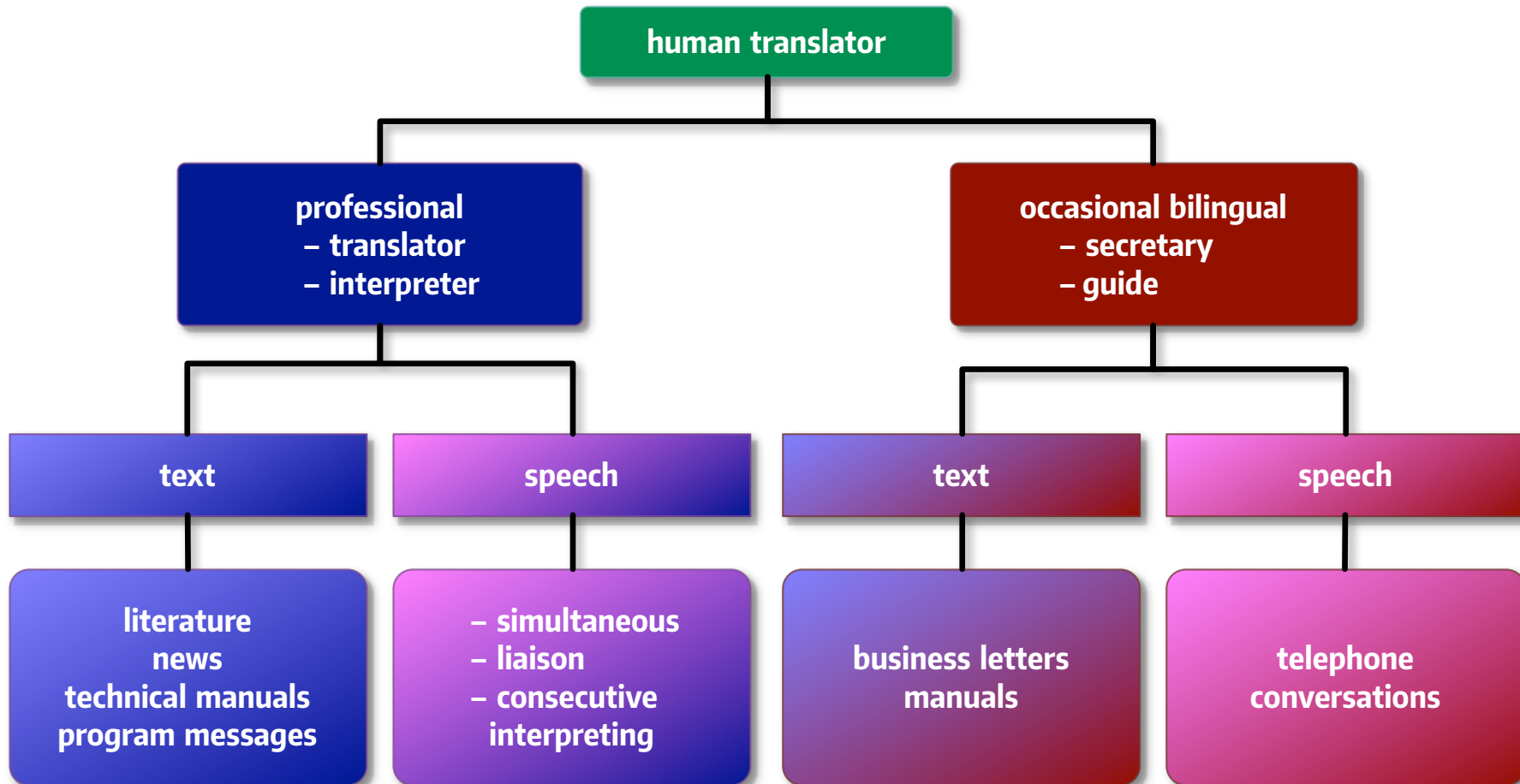
INTRODUCTION

What is this session about?

- A broad overview of Machine Translation (MT)

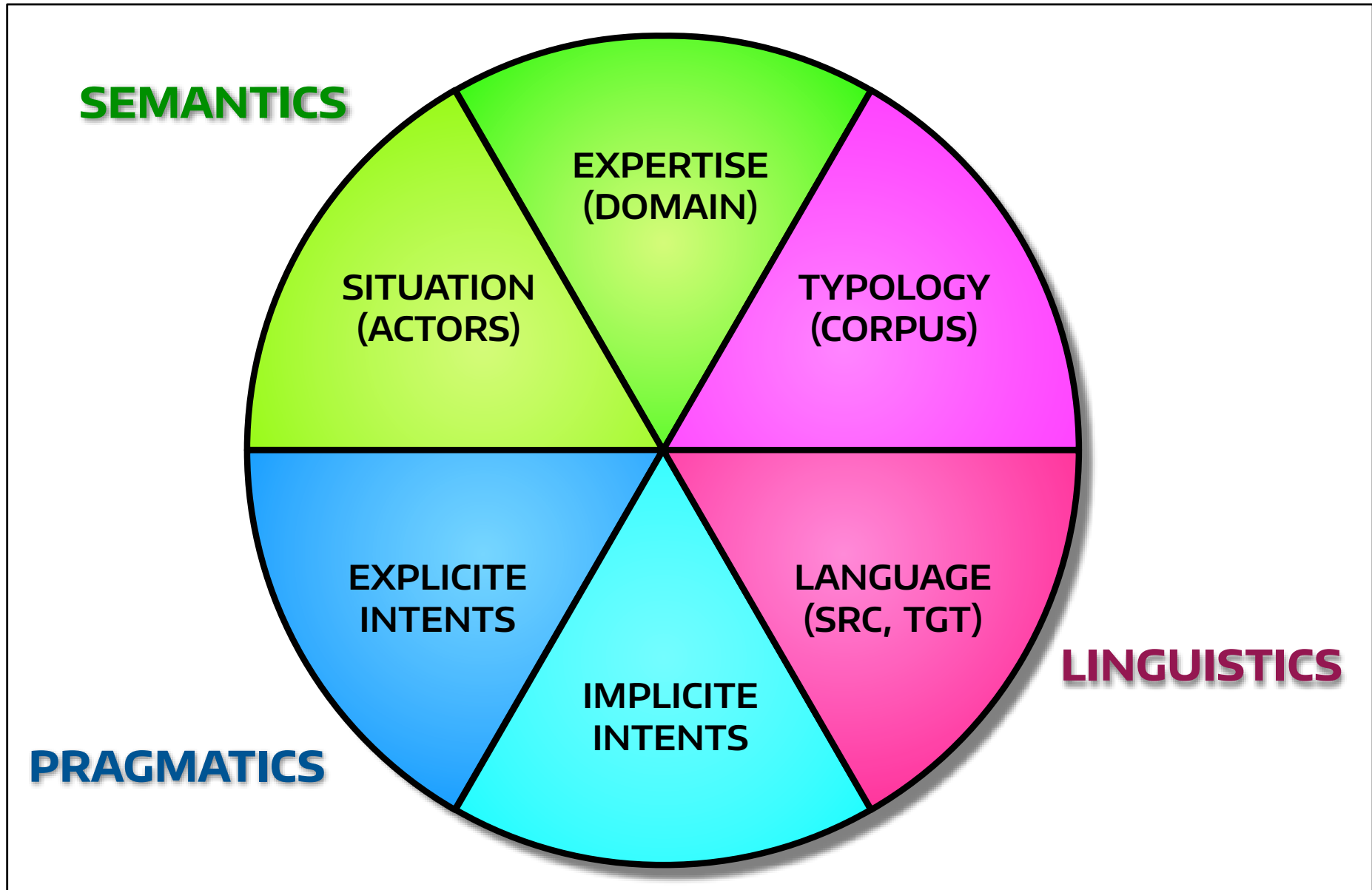- Vocabulary

- Techniques

- Evaluation
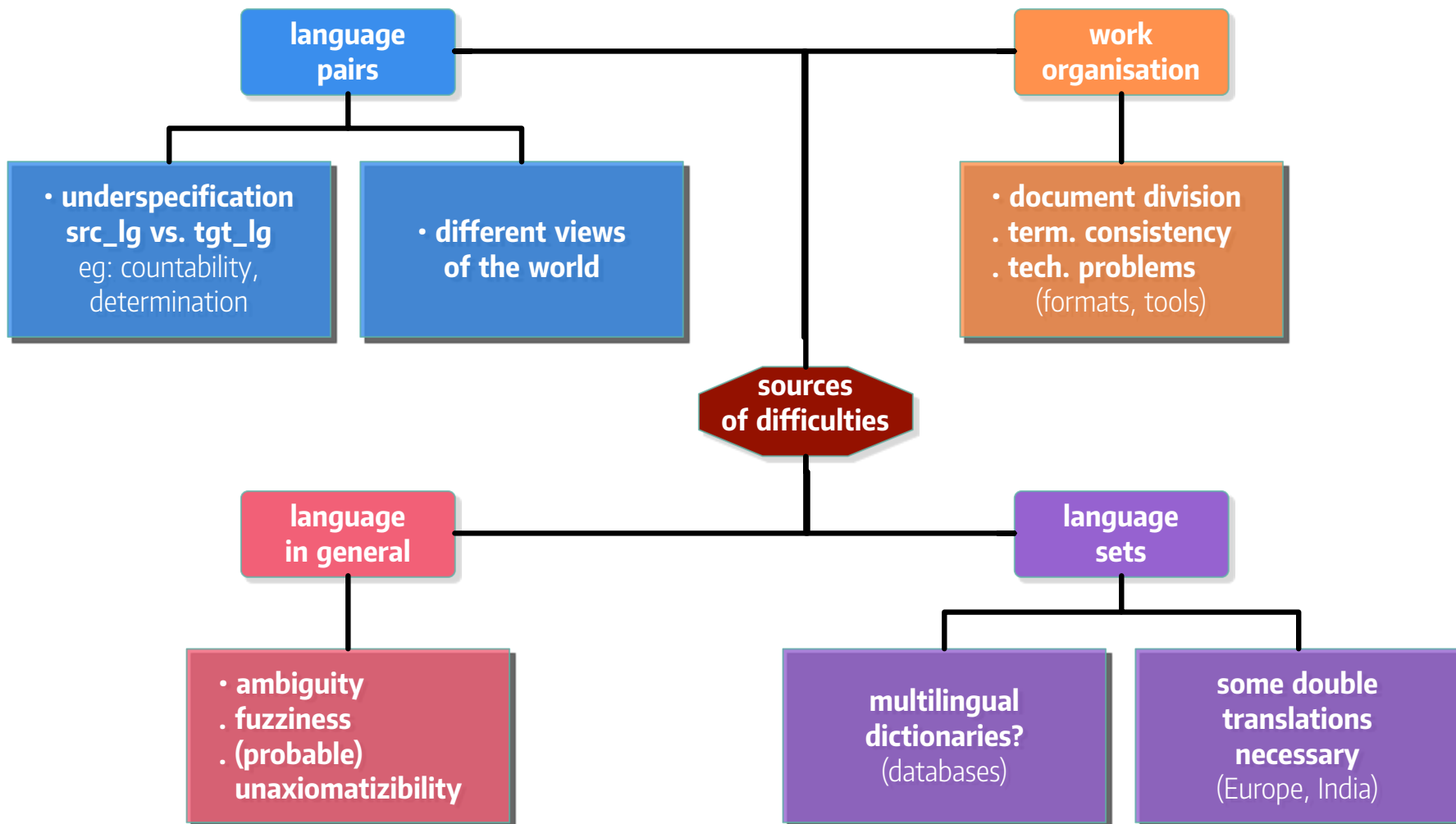
HUMAN TRANSLATION

Human Translation is multiple

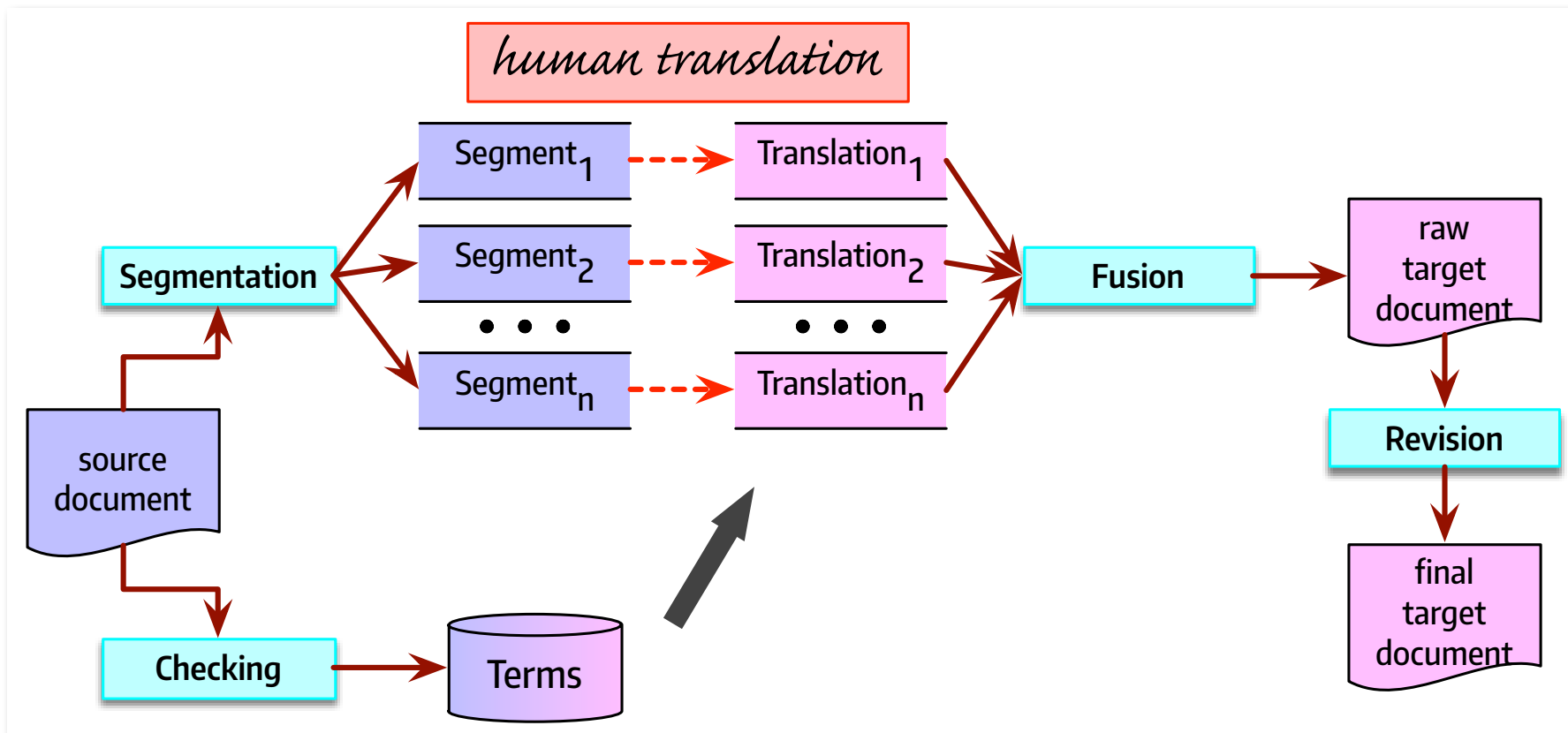

```
                        human translator
           ┌──────────────────┴──────────────────┐
    professional                          occasional bilingual
    – translator                          – secretary
    – interpreter                         – guide
   ┌──────┴──────┐                       ┌──────┴──────┐
 text          speech                  text          speech
   │             │                       │             │
literature   – simultaneous         business letters  telephone
news         – liaison              manuals           conversations
technical    – consecutive
manuals        interpreting
program
messages
```

Human Translation Knowledge Sources

SEMANTICS

EXPERTISE (DOMAIN)

SITUATION (ACTORS)

TYPOLOGY (CORPUS)

EXPLICITE INTENTS

IMPLICITE INTENTS

LANGUAGE (SRC, TGT)

LINGUISTICS

PRAGMATICS

Human Translation is difficult!

D'après Christian Boitet

Human Translation Workflow

MACHINE TRANSLATION

Nature of MT

- Machine Translation
  - not a science, nor a set of recipes
  - it is a
  - scientific technology
- Machine Translation
  - a set of methods that are progressing through
    - integration of new ideas
    - incremental improvement of know-how

A very difficult problem by nature

- It difficult for humans!

- + Fuzziness: a natural language can not be completely axiomatized (precisely modelled)
  - any formal description will be either sub-generating or over-generating
  - only "small" or controlled sub-languages can be modeled

- + Huge sizes
  - millions of concepts/terms
  - variety of grammatical phenomena and their interactions

HT and MT are different

- MT is not there to replace humans
  - cars are not horses!
- MT does not mimic humans
  - planes do not flap their wings!

Inherent Limit to the Automation Problem

Coverage ✖ Automation ✖ Quality << 1

- improvement of one of the criteria is necessarily to the detriment of the other two
  - ↗ coverage ⟹ ↘ automation and/or quality
  - ↗ automation ⟹ ↘ coverage and/or quality
  - ↗ quality ⟹ ↘ coverage and/or automation

- Exemples
  - C x A ≈ 100% ⟹ translation available on the Web
    - *Google, Systran, …*
  - A x Q ≈ 100% ⟹ specialized systems
    - *Meteo, AltFlash*
  - C x Q ≈ 100% ⟹ Dialogue-Based MT
    - *Jets, Catalyst, Lidia*

Who is the user?

Machine Translation Workflow

D'après Christian Boitet
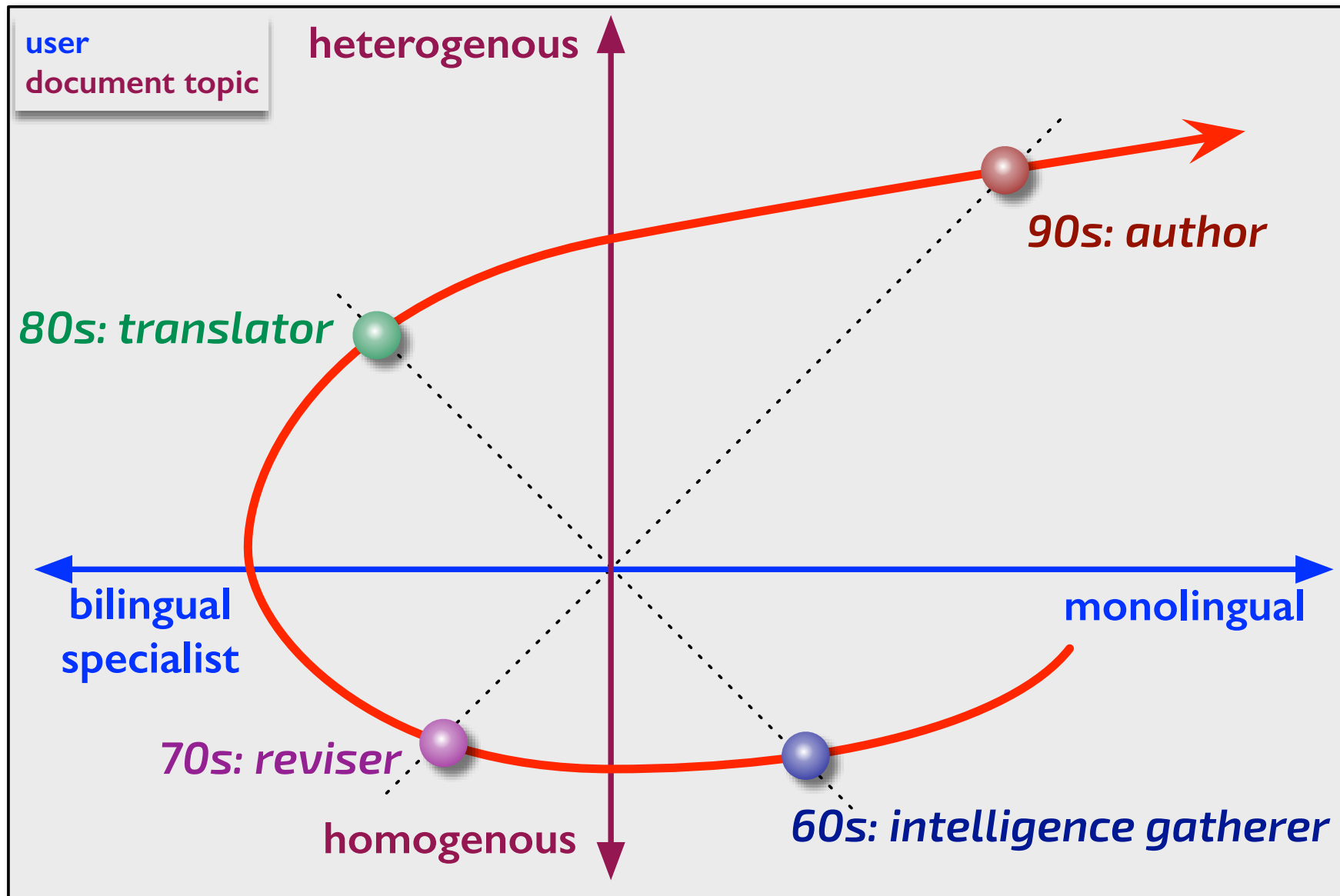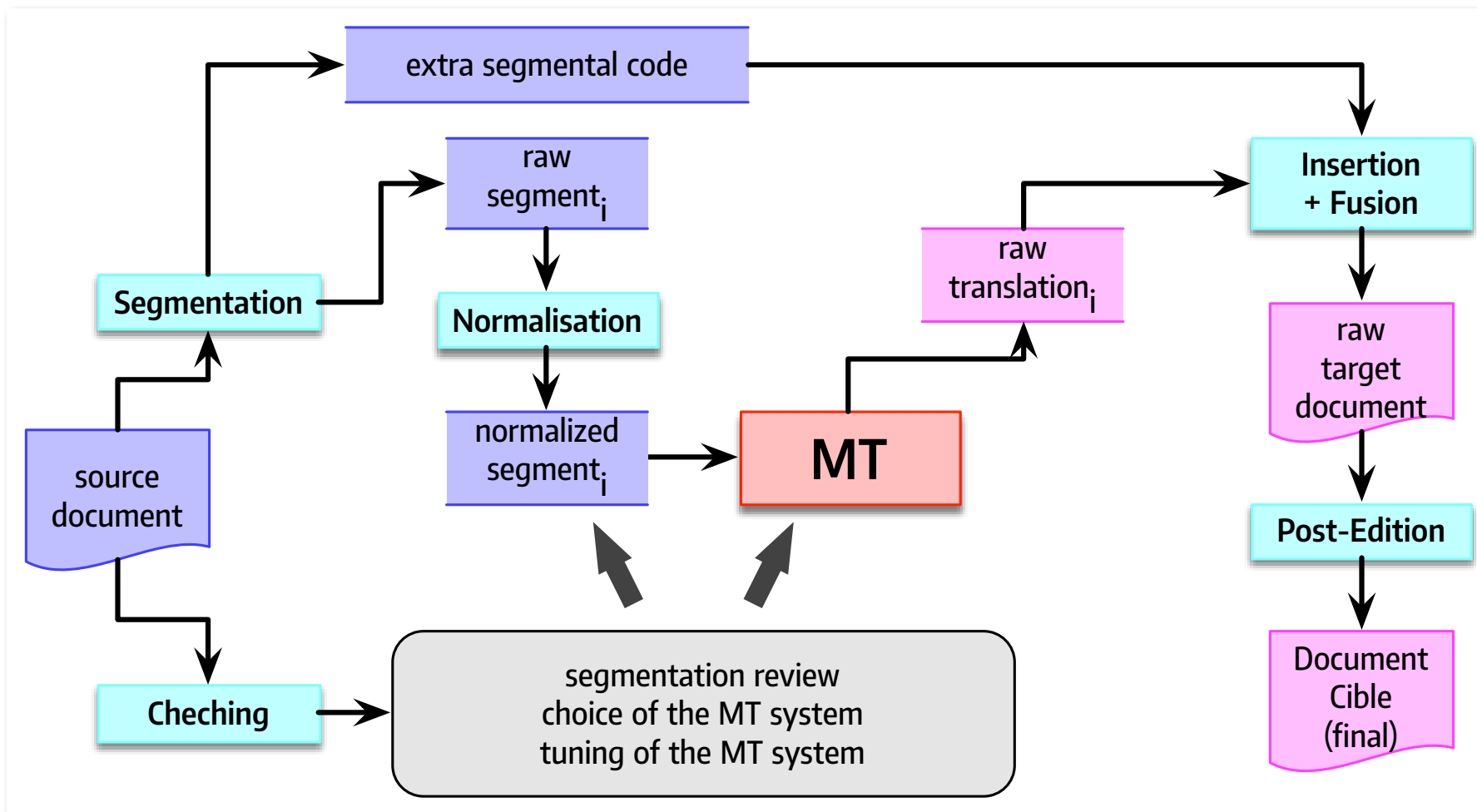
TRANSLATION NICHES

Niches

Re-creation

- advertisement, poetry, essay, novel

Localization

- adapt a content to a given cultural environment (e.g. a website, a navigation system, a software, …)

Localization

- LISA definition (Localization Industry Standards Association)
  - "modify products or services by taking into account differences in distinct markets"
- 3 kinds of problems
  - linguistics
    - translation of the user interface
    - translation of the manuals
  - content and culture
    - information presentation (icons, graphics, colors, communication, …)
    - examples: – interface of software in a car: driving on the left or on the right; – direction of writing and reading (left-to-right, right-to-left)
  - technical
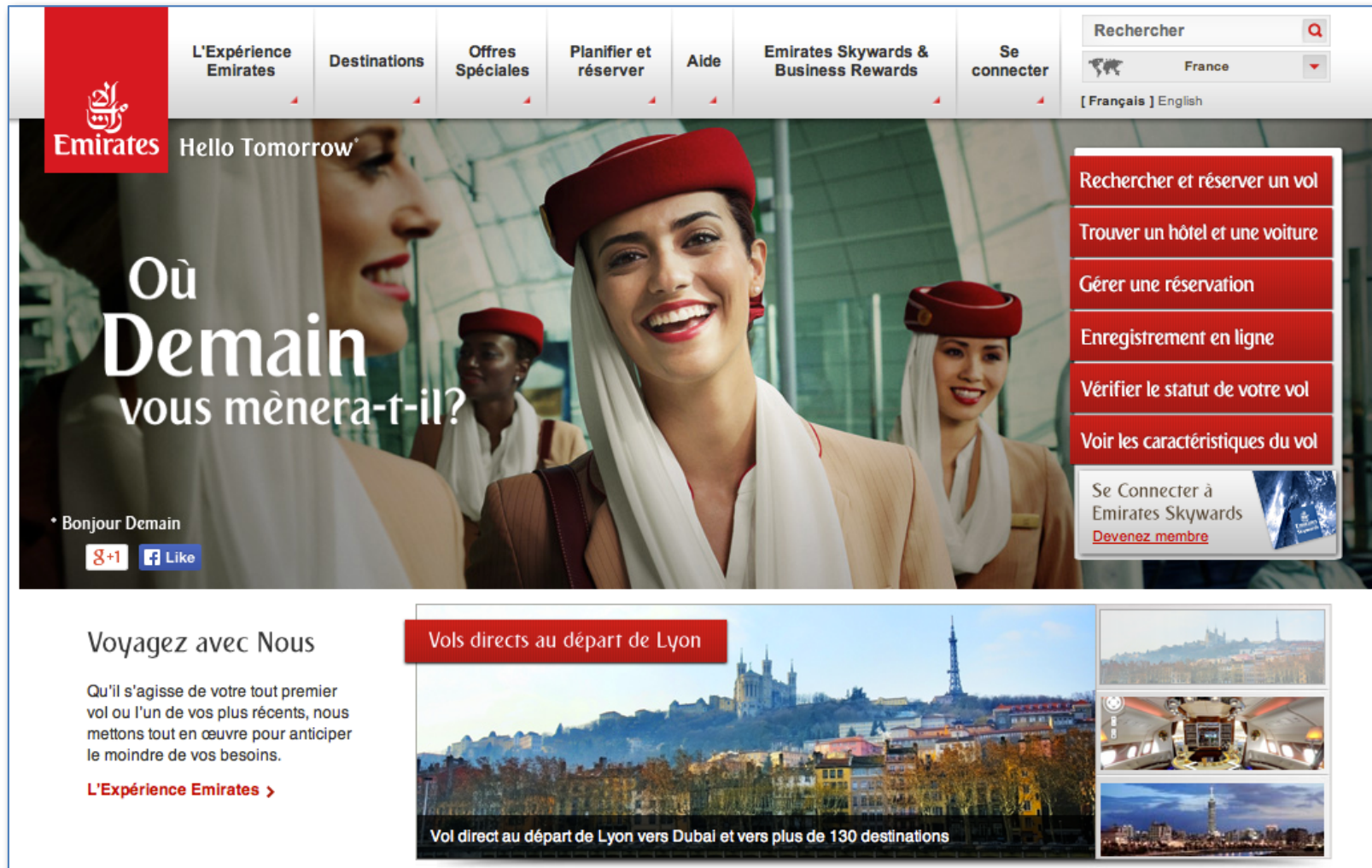   - support of non roman scripts (Arabic is bi-directional: – numbers and foreign words from left to right, – other from right to left)
    - character encoding (1, 2, 4 bytes)

Website Localization (1/6)

 French version of the Website

Website Localization(2/6)

French version of the Website

localization: dates and calendar

Website Localization(3/6)

Japanese version of the Website

Website Localization (4/6)

Japanese version of the Website

localization: dates and calendar

Website Localization (5/6)

Arabic version of the Website

reading from right to left (logo → right, menu → left)

Website Localization (6/6)

Arabic version of the Website

localization: dates and calendar

Niches

Re-creation

- advertisement, poetry, essay, novel

Localization

- adapt a content to a give cultural environment
  (e.g. a website, a navigation system, a software, …)

Diffusion (Dissemination, high quality required)

- technical documentation whose content must be strictly rendered, without addition or omission even if the style may feels "unnatural"
- a bad translation can have disastrous repercussions

Same pictures for all the versions

clear picture of the original English

Creating events

1 Launch the **Calendar** app and tap ⊕ or double-tap a date.

If the date already has saved events or tasks in it, tap the date and tap ⊕.

2 Enter event details.

Set a start and end date for the event.

Enter a title. — Title

Change the event's colour.

All day

Start   Thu, 12 Jan 17:00
End   Thu, 12 Jan 18:00

Set an alarm. — • My calendars

Select a calendar to use or sync with.

10 mins before   ADD

Enter the location of the event. — Location   MAP

Add a note. — Notes

Attach a map showing the location of the event.

Repeat
Never

Add more details. — Time zone

3 Tap **SAVE** to save the event.

Same pictures for all the versions

blurred picture of the original English

Créer des événements

1 Démarrez l'application **Calendrier**, et appuyez sur ⊕ ou appuyez deux fois sur une date.

Si la date comporte déjà des événements ou tâches, appuyez sur la date, puis sur ⊕.

2 Accédez aux détails de l'événement.

Saisir un titre

Définir une alarme

Indiquer l'endroit où se déroule l'événement

Ajouter une note

Ajouter des détails

Définir une date de début et de fin pour l'événement

Changer la couleur de l'événement

Sélectionner un calendrier à utiliser ou à synchroniser

Joindre une carte indiquant le lieu de l'événement
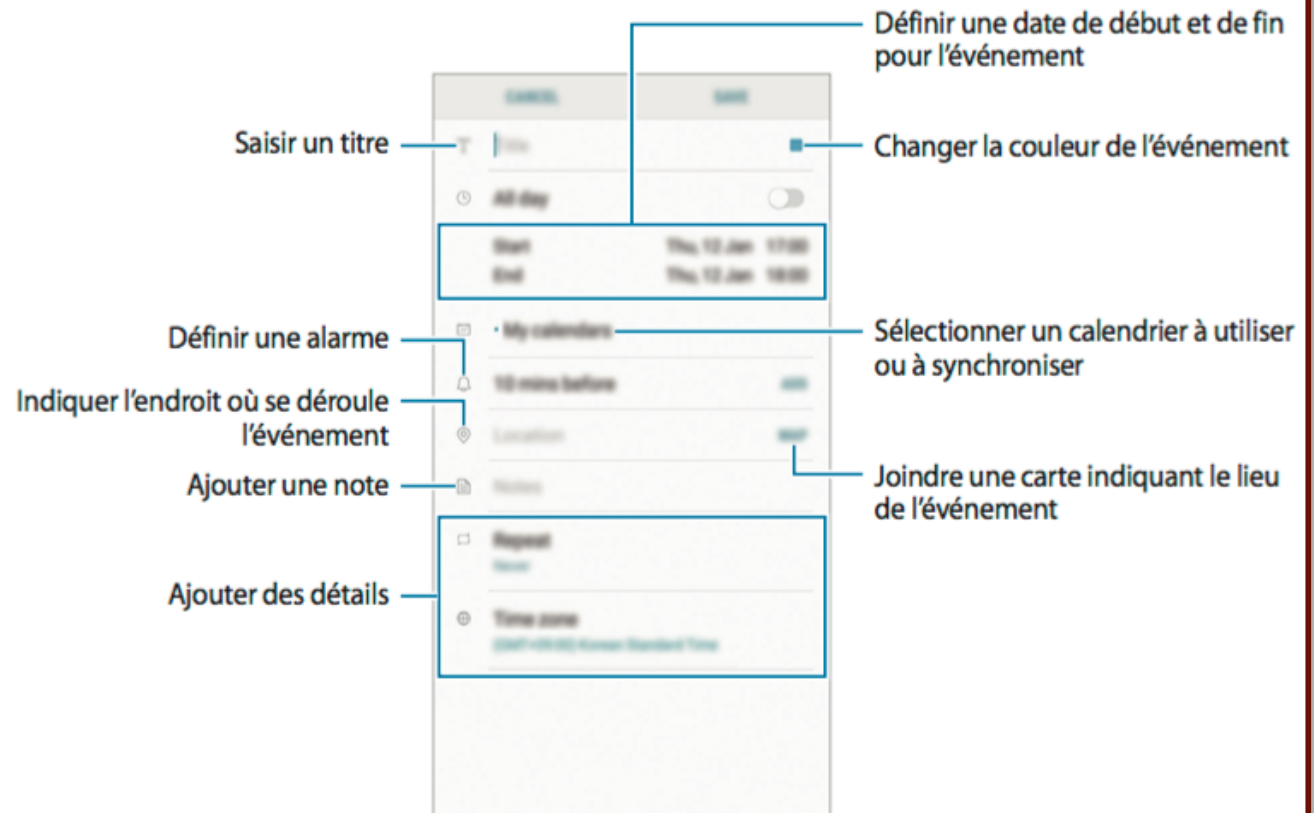
3 Appuyez sur **ENREGISTRER** pour enregistrer l'événement.

Same pictures for all the versions

blurred
picture
of the
original
English
with
horizontal
flip

إنشاء أحداث

١ قم بتشغيل تطبيق **التقويم** واضغط على ➕ أو اضغط ضغطًا مزدوجًا على التاريخ.

إذا كان التاريخ به أحداث أو مهام محفوظة، فاضغط على التاريخ واضغط على ➕.

٢ أدخل إلى تفاصيل الأحداث.

ضبط تاريخ لبداية الحدث ونهايته.

تغيير لون الحدث.

إدخال عنواناً.

تحديد تقويمًا لاستخدامه أو للمزامنة معه.

ضبط المنبه.

إدخال موقع الحدث.

إرفاق خريطة تعرض موقع الحدث.

إضافة ملاحظة.

إضافة المزيد من التفاصيل.

٣ اضغط على **حفظ** لحفظ الحدث.

Niches

Re-creation

- advertisement, poetry, essay, novel

Localization

- adapt a content to a give cultural environment
(e.g. a website, a navigation system, a software, …)

Diffusion (Dissemination, high quality required)

- technical documentation whose content must be strictly rendered, without addition or omission even if the style may feels "unnatural"
- a bad translation can have disastrous repercussions

Acquisition (Assimilation)

- "screening" of written texts and simultaneous interpretation

Original here

▼飛鳥・奈良時代（着物に似た、右前の衣服）

飛鳥・奈良時代には遣隋使（けんずいし）や遣唐使（けんとうし）などによりさまざまな分野で中国のものが取り入れられました。

飛鳥時代には聖徳太子により冠位十二階が制定され、官吏の位階を十二階に分け、位により冠と衣服の色が定められ、また奈良時代には礼服（らいふく）、朝服（ちょうふく）、制服（せいふく）を位により服装を三分類する、三公服が制定されました。

衣服では、衿（えり）を立てたコート状のもので袍（ほう）形式といわれるものが支配者階級の服装（朝服）として男子は衣（きぬ）に袴（はかま）、女子は衣に裙（も）というものを着ていたようです。また褶（ひらみ）というものを裳や袴の上からつけていたようです。奈良時代には今までの左前の打ちあわせから、現在の「右前」の打ち合わせに改められたようです。

※袍形式（ほうけいしき）
装束を構成する前開きのガウン状の表着（うわぎ）のことを袍といい、それを着用した形式。

▼ Asuka · Nara era (kimono resembles, right front garment)

In the Asuka and Nara period, things of China were introduced in various fields by the enforcers' squirrels (Kenzuji) and envoys (Kenshi).

In the Asuka period, the crown of the twelfth floor was enacted by Prince Shotoku, the ceremonial position of the civil service officer was divided on the 12th floor, the colors of the crown and the clothes were determined by the position, and in the Nara period, the clothes (rafu), morning uniform Sanko clothing was enacted, classifying three clothes according to the position of the uniform (school uniform).

In clothes, a coat shaped collar (eiri) that is said to be in the form of a bamboo clothing (morning clothes) as men's clothing (morning clothes) as men's clothes (kimono) hakama (girls), girls in clothing It seems that she was wearing something like that. Moreover, it seems that I put a thing called a fold (Hirami) from the top of the sushi and hakama. In the Nara era, it seems that from the meeting of the left front until now, it was changed to the present "right front" meeting.

※ 袍 form (Housing)
A garment-like garment (huge) that opens before opening that constitutes a coin is called,, and it is a form that wears it.

French from google translate (2017/06/26)

▼ Asuka et Nara période (semblable au kimono, vêtements de

La Asuka et l'ère Nara Chine des choses a été incorporée dans une variété de domaines tels que par Yazuishi (Kenzuishi) et envoyés (des envoyés).

La période Asuka adopté douze Level Cap et le système Rank est par le prince Shotoku, le rang des fonctionnaires divisé en douze étage, la couleur de la couronne et des vêtements est déterminée par la position, aussi dans la robe de la période Nara (revenir), les vêtements du matin (Chofuku), uniformes (uniformes) à trois positions classement habillé par les trois ministres des vêtements ducales a été adoptée.

Dans les vêtements, 裙 comme col Hou dans ce manteau comme ça fait un (Eri) (loi) classe dirigeante de la robe est ce qu'on appelle une forme (dans les vêtements du matin) le hakama des hommes aux vêtements (soie) (hakama), les filles dans les vêtements (aussi), il semble avoir été vêtu d'une chose appelée. En outre, il semble avoir la chose appelée 褶 (Hirami) du haut de Sang et hakama. De la réunion de l'avant gauche jusqu'à la période Nara maintenant, il semble a été changé à la réunion du « droit devant » actuel.

※ Hou le format (format de la loi)
Beaux vêtements de table comme robe avant d'ouvrir qui composent les costumes de (veste) avec borique, le format que vous portez.

Japanese Web Page about Kimono

English from google translate (2017/06/26)

▼ **Asuka e il periodo Nara (simile al kimono, abiti di destra)**

L'era Asuka e Nara la Cina di cose è stato incorporato in una varietà di campi come da Yazuishi (Kenzuishi) e gli inviati (inviati).

Il periodo di Asuka emanato Dodici Livello Cap e Rank System è dal principe Shotoku, il rango dei funzionari diviso in dodici piano, il colore della corona e l'abbigliamento è determinata dalla posizione, anche nel periodo Nara veste (ritorno), abbigliamento mattina (Chofuku), uniformi (posizione divise) a tre classificazione vestita dal, tre vestiti ministri ducali è stata emanata.

In abiti, 裙 come collare Hou in quello cappotto del genere ha fatto un (Eri) (legge) classe dirigente del vestito è quello che viene definito come una forma (nei vestiti mattina) hakama uomini di abbigliamento (seta) (hakama), ragazze in vestiti (anche) che sembra essere stato indossare una cosa chiamata. Inoltre sembra avere con la cosa chiamata 褶 (Hirami) dalla parte superiore del Sang e hakama. Dalla riunione della parte anteriore sinistra fino al periodo Nara ora, sembra stato cambiato alla riunione della corrente "anteriore destro".

※ Hou formato (formato legge)
tovaglie abito simile a Nizza prima di aprire che compongono i costumi che di (giacca) con borico, formato che si indossa.

HOW TO DESCRIBE A MT SYSTEM?

Description Grid

Operational Architecture	development and use of the system

Linguistic Architecture
 	
    Objects
 	
    see Vauquois' triangle
	intermediate representations
	
	    direct, semi-direct,
	
	    transfer (≥ 7 variants)
	
	        2 lexical spaces
	
	    IL (≥ 2 variants)
	
	        3 lexical spaces

Computational Architecture
 	
    automatic processes
 	
    human interaction, if any
	programming paradigms
	
	    direct programming
	
	    RBMT (rules, dictionaries, automata…)
	
	empirical (corpus-based)
	
	    SMT, PSMT (unsupervised)
	
	    EBMT (≥ 3 variantes)
	
	        ± supervised
	
	hybrid

Operational Architecture

development and use of the system

Operational Architecture

- Conditions for the development and use of a system
  - Tasks and users *(to be continued)*
  - Language pair(s), volumes, genres *(to be continued)*
  - Possible human intervention(s) *(to be continued)*
  - Resources available
    - Data
      - Large parallel corpus (annotated or not) for empirical MT
    - Human
      - Computational linguists, lexicographers for expert MT

OA: Tasks and Users

- Help bilingual users produce good translations
  - diffusion

- Help users understand a non-(or little) known language
  - acquisition (translation testing, assimilation)

- Assist in interpersonal communication (chat, speech translation)

OA: Language pair(s), volumes, genres

Translation paths or directions

- 1 → 1 :
  - ALT/JE (NTT)
- 1 ↔ 1 :
  - ATLAS-II (fujitsu) : dicos 29 M words + 5,57 M words
  - Converser for Healthcare (Spoken Translation)
- 1 → N :
  - MedSLT
- 1 ↔ N :
  - Phraselator (US Army, speech)
- N ↔ N :
  - for multilingual debate, chats; multilingual peacekeeping forces

OA: Possible Human Intervention(s)

0% (100% automated)   100% (human)

(FAHQT)
Fully Automatic
Hight Quality
Translation

(HAMT)
Human-Aided
Machine
Translation

(MAHT)
Machine-Aided
Human
Translation

Traditional
Human
Translation

(CAT)
Computer Assisted Translation
(TAO)
Traduction assistée par ordinateur
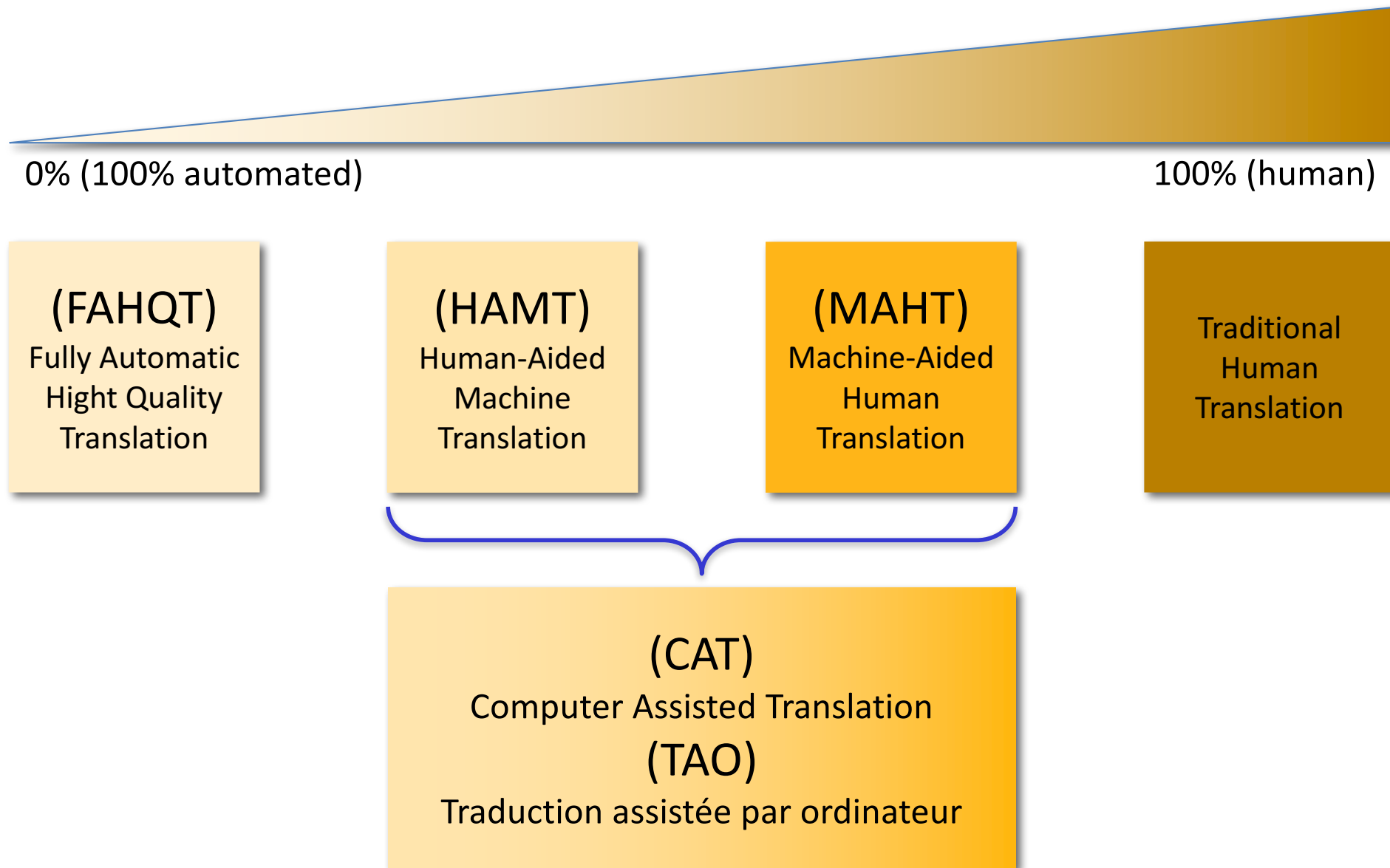
FAHQT

Fully Automatic High Quality Translation

Possible in restricted settings

- the context must be very well defined
- these are normally autonomous systems

FAHQT: examples

MÉTÉO

- operational since 1977

- translation of Canadian weather forecast

- English ↔ French

- 60000 words/day ; corrections: 3%

ALTFLASH

- operational since 2001

- translation of Nikkei Stock Exchange newsflashes

- Japanese → English

- raw results without human revision, because "pro" quality is less crucial than for the weather

HAMT

Human-Aided Machine Translation

- User involved in the translation process
  - context must be well defined
  - interactive systems
- Three possible approaches
  - pre-edition
  - post-edition (revision)
  - interaction during the translation itself

HAMT with Pre-Edition

Pre-Edition for what purpose?

source document → pre-edition → pre-edited document → autonomous MT system → target document
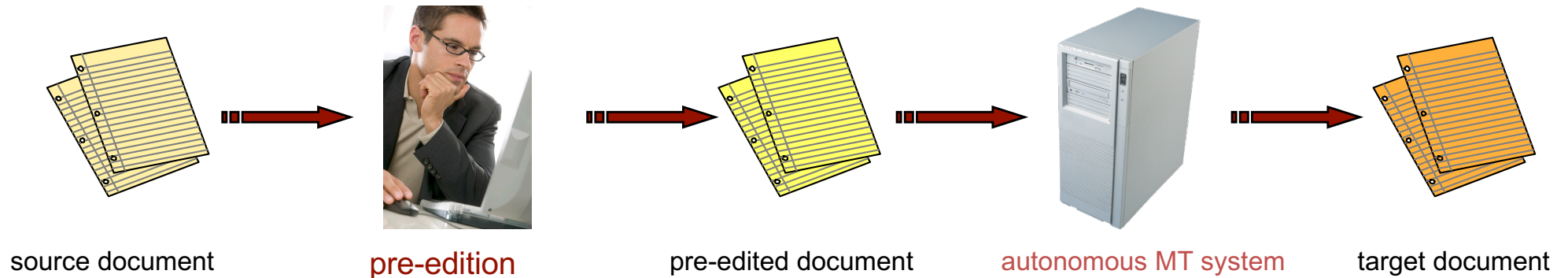
Search for problems that the system will encounter and remove them

- identification of proper names
- choice of categories for homographs
- indication of embedded clauses
- bracketing of coordinated structures
- reformulate the text in a controlled language (extreme)

HAMT with Post-Edition

 Post-Edition for what purpose?

| source document | autonomous system | raw translation | post-edition | target document |
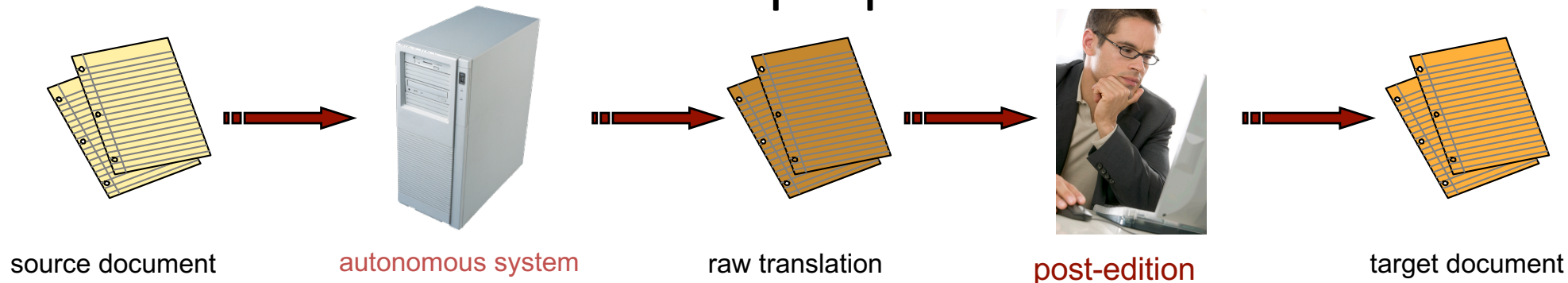
 Correct the raw translation produced by the system according to the quality to be obtained:

- very little if it's information retrieval
- many and by a specialist familiar with the field if they are texts that must be published and disseminated

 Almost all operational solutions

HAMT with Interaction

Interaction for what purpose?

document source   système interactif   document cible

- assist the system during the translation process

interaction for whom?

- a translator: interaction for specialist
- a monolingual user: interaction for non specialist
  - interaction during redaction step
  - interaction during translation step
  - no interaction with the translation output

MAHT

Machine Aided Human Translation

- The user is assisted in the translation process

- Tools exist
  - checkers
    - spelling, grammar, style
  - dictionaries
    - monolingual, bilingual, & encyclopedic
  - translation memories
    - linguee

- Idea: Integrate services into a
  Translator's Workstation

Linguistic Architecture

Pyramid diagram with levels on the left:
- Deep Understanding level
- Interlingua level
- Logics-Semantic level
- Mixing levels
- Syntaxico-functional level
- Syntagmatic level
- Morpho-Syntactic level
- Graphemic level

Transfer labels (left to right on pyramid):
- conceptual transfer
- Semantic transfer
- Multilevel transfer
- Syntactic transfer (deep)
- Syntactic transfer (surface)
- Semi-direct
- Direct

Right side structures:
- Ontological interlingua
- Semantico-lingustic interlingua
- SPA-structures (semantic & predicate-argument)
- Multilevel description
- F-structures (functional)
- C-structures (constituent)
- Tagged text
- Text

Linguistic Architecture

Three options

- 1 step translation (direct translation)
- 3 steps translation
  - analysis
  - transfer
  - generation
- 2 steps translation (Translation with a abstract pivot)
  - analysis (into the pivot)
  - generation (from the pivot)

Direct Translation

- The system is a black box (1 step)
  - succession of treatments
    - from the source language
    - to the target language
    - without producing any intermediate structure

Source Language text	→	DIRECT TRANSLATION ENGINE	→	Target Language text

Analysis, Transfer & Generation

The system contains 3 black boxes (3 steps)

SL Text REPRESENTATION →	**Transfert** →	**SL Text REPRESENTATION**
↑ **Analysis**	**TRANSFERT-BASED TRANSLATION ENGINE**	↓ **Generation**
↑ **Source Language text**		↓ **Target Language text**

analysis: a data structure representing the source language (source knowledge)

transfer: a data structure representing source language into the target language (source & target knowledge)

generation: a text in the target language (target knowledge)

Translation with Pivot

The system contains 2 black boxes (2 steps)

SL Text MEANING REPRESENTATION

Analysis

PIVOT-BASED TRANSLATION ENGINE

Generation

Source Language text

analysis: a data structure representing the "meaning" of the source language (source, domain, world knowledge)

generation: a text in the target language (target knowledge)

Target Language text
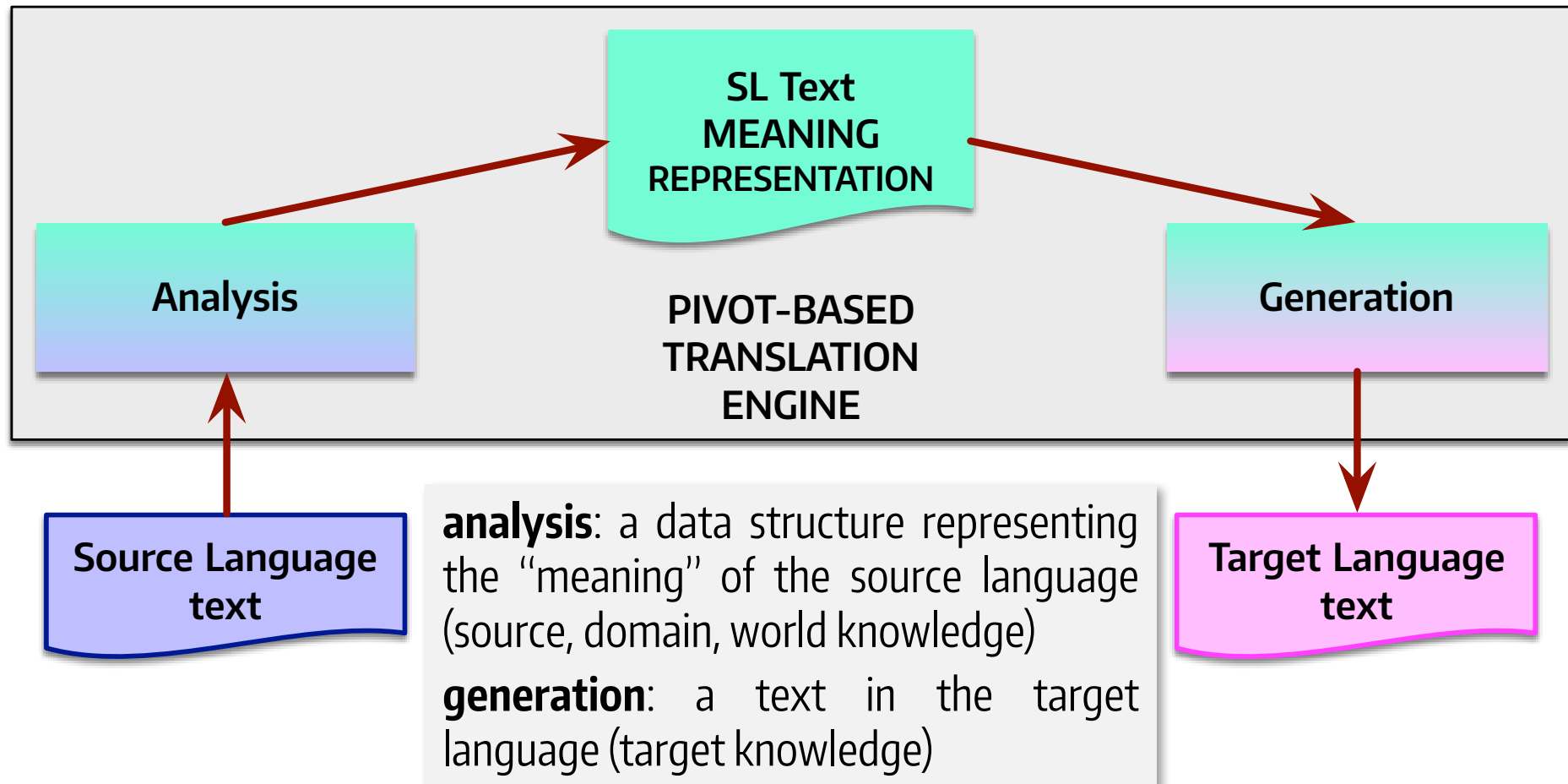
Vauquois' Triangle
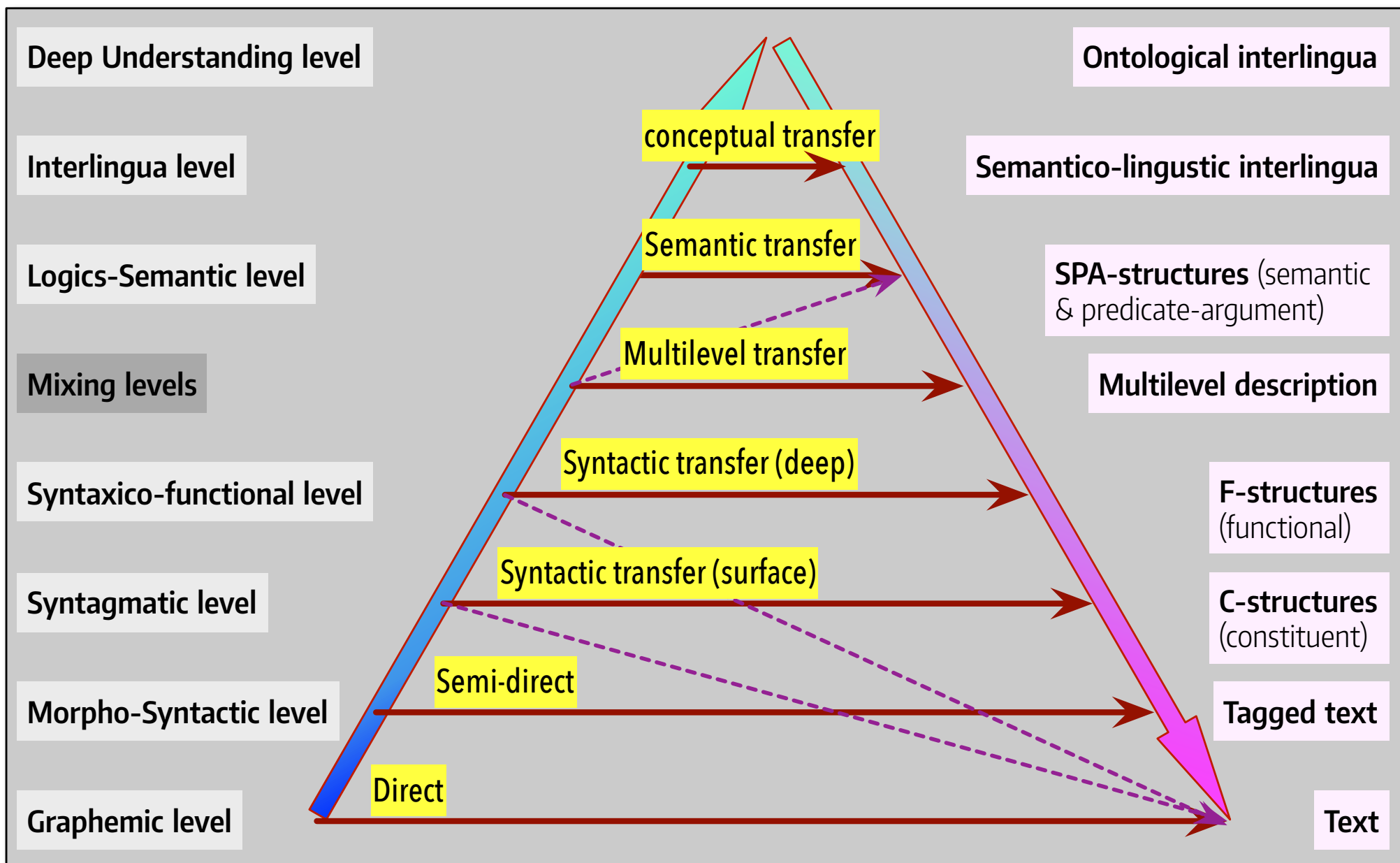
Examples of Structures (1/3)

🟥 Multilevel Concrete Syntagmatic Tree

🔷 Le capitaine a rapporté un vase de Chine

occurrence
Lexical Unit
Syntactic or Syntagmatic Class
Syntactic Function
Logic and Semantic Relation

'PHVB'
PHVB

'DGN'
GN
SUJ
ARG0

'NV'
PHVB

'DGN'
GN
OBJ1
ARG1

'GP'
GN
CIRC
UNDE

le	capitaine	a	rapporté	un	vase	de	Chine
'le'	'capitaine'	'avoir'	'rapporter'	'un'	'vase'	'de'	'Chine'
D	N	V	V	D	N	S	N
DES	GOV	AUX	GOV	DES	GOV	REG	GOV
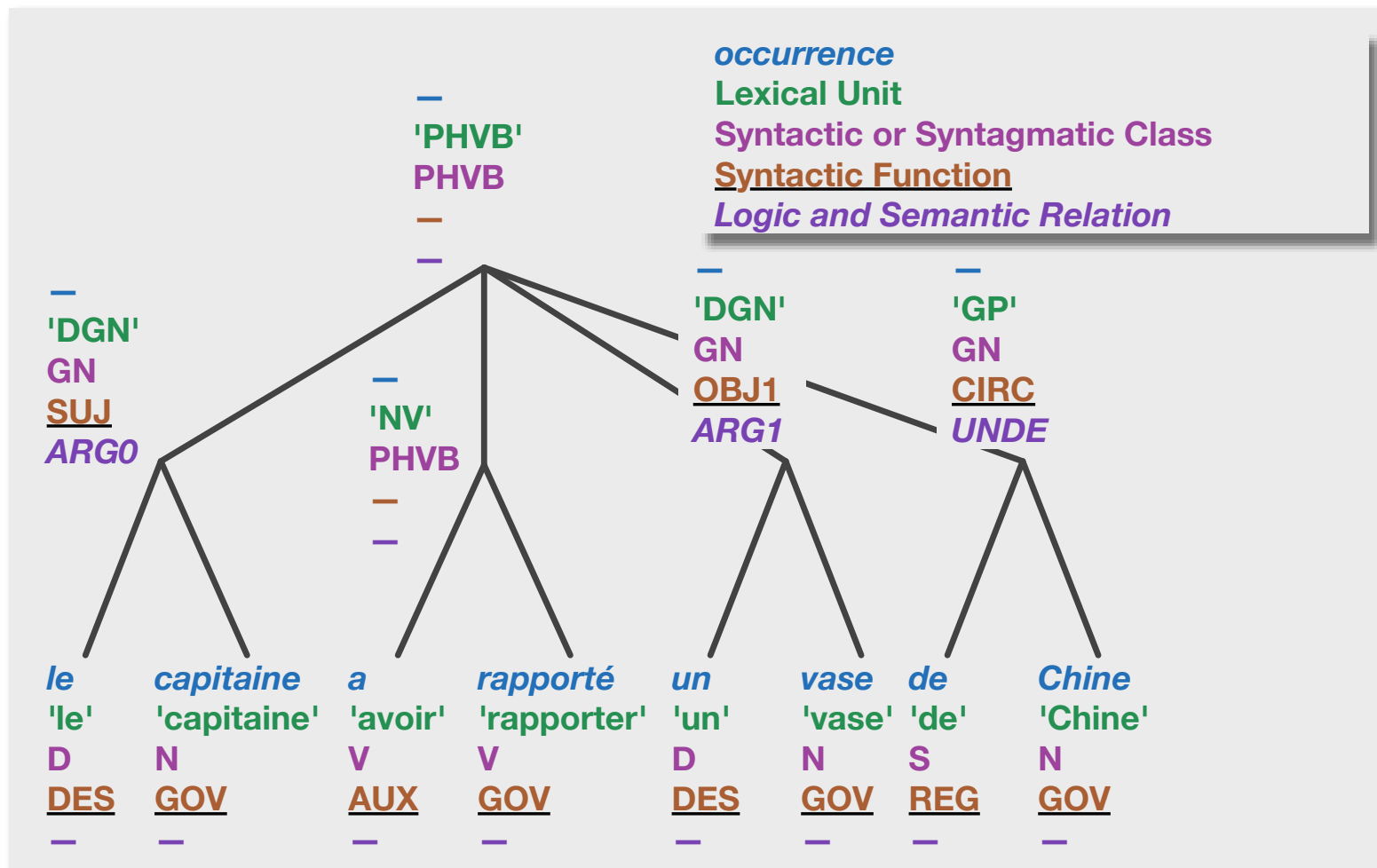
Examples of Structures (2/3)

- Semantico-Pragmatic Task-Related Pivot (ontological)

  - \<client> et je voudrais une chambre du dix au quinze septembre à trento

  > {c:give-information+disposition+room (
  >   conjunction=discourse,
  >   disposition=(desire, who=i),
  >   room-spec=(identifiability=no, quantity=1, bedroom),
  >   location=name-trento,
  >   time=(start-time=(md=10), end-time=(md=15, month=9)))}

  - speech act: **give-information**
  - concepts : **disposition**, **room**
  - dialog act: **give-information+disposition+room**
  - arguments:
    - **disposition=(desire, who=i)**
    - **room-spec=(identifiability=no, quantity=1, bedroom, location=name-trento)**
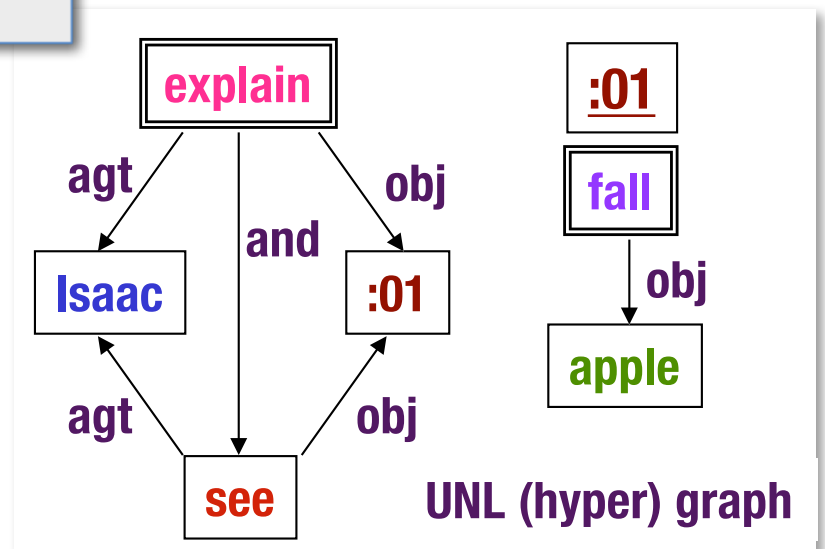   - **time=(start-time=(md=10), end-time=(md=15, month=9))**

Exemples de structures (3/3)

Semantico-Linguistic Pivot (UNL)

Isaac sees that an apple falls and he explains it.

```
agt(explain(icl>do).@entry,Isaac(icl>proper noun))
obj(explain(icl>do).@entry,:01)
obj:01(fall(icl>occur).@entry,apple)
and(explain(icl>do).@entry,see(icl>do))
agt(see(icl>do),Isaac(icl>proper noun))
obj(see(icl>do),:01)
```


UNL (hyper) graph

Computational Architecture

Programming Paradigms

Empirical methods

Hybrid Approaches

Computational Architecture

- The method of preparing each of the phases constituting the steps …
  - analysis: graphemic, morphological, syntactic, …
- … this dimension is in fact orthogonal to the linguistic architecture
- Three families
  - Expert Approaches
  - Empirical Approaches
  - Hybrid approaches

Expert Approaches

- Direct approach, procedural
  - "Lingual craftsmanship"
  - grammars, automata and dictionaries are "coded" directly in a programming language

- Rules-based approach
  - Inspired by formal linguistics and compilation
    - Formal grammars, automata and dictionaries

      +

    - Engine

Empirical Approaches

Machine learning

- supervised or not

- of correspondences from more or less annotated data
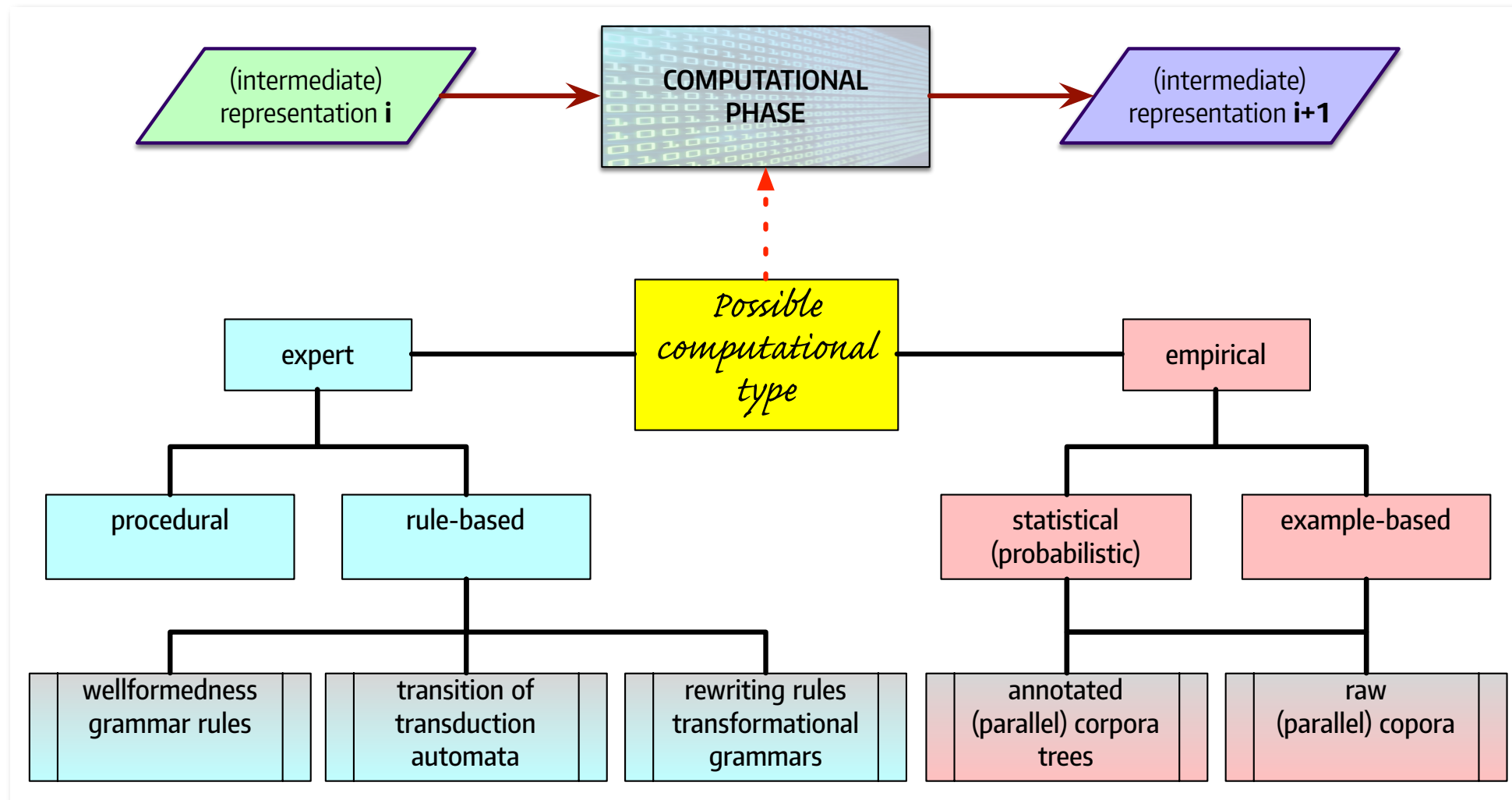
  - text → text

  - text → structure

  - structure → structure

  - structure → text

Computational Architecture

Summary of the possible approaches

D'après Christian Boitet

CONCLUSION

First Remarks

- The translating function can only be automated for assimilation and dissemination

- Near perfect TAO is possible in restricted settings: Weather
  - Technique not extensible for the moment
  - Examples isolated, few real situations

- Current MT toolboxes and/or "off the shelf systems" adapted to the documents to be translated makes it possible to obtain usable results of sufficient quality
  - A fare number of toolboxes
  - A multitude of systems

Sketch of a draft!

REFERENCES

Further readings *(old but still very informative)*

- The State of Machine Translation in Europe and Future Prospects (2002)
  - http://ccl.pku.edu.cn/doubtfire/NLP/Machine_Translation/Overview/Article%20-%20MT_John_Hutchins.htm
- Machine translation: An Introductory Guide (1994)
  - http://www.essex.ac.uk/linguistics/external/clmt/MTbook/
- Survey of the State of the Art in Human Language Technology (1996)
  - http://www.dfki.de/~hansu/HLT-Survey.pdf
- La traduction automatique (2001)
  - http://www.etudes-francaises.net/entretiens/15traduction.htm
- Machine translation: Publications by John Hutchins
  - http://www.hutchinsweb.me.uk/

Web Pointers

Free/open-source machine translation software

- Mikel L. Forcada
- http://fosmt.org

Statistical MT

- everything you need to know, almost every toolbox
  - http://www.statmt.org/
- Thot

Neural MT (NMT)

- Rico Sennrich, Barry Haddow (2017) *Practical Neural Machine Translation*. EACL-2017 Tutorial
- OpenNMT
- NEMATUS