European Commission

European Commission Machine Translation

Hybrid Approaches to Machine Translation in MT@EC and CEF.AT

What is behind the EC's MT service and how do we want it to evolve?

Andreas Eisele
Project Manager - Machine Translation Engines
Directorate-General for Translation R3.2

WS of the European Language Resource Coordination
Atelier Traduction Automatique en Luxembourg, 14.6.2016

Hybrid Approaches to Machine Translation in MT@EC and CEF.AT

- What are MT@EC and CEF.AT?
- Our users and some key requirements
- Architecture: From pure Statistical to Hybrid MT
- How to improve the translation quality
  - Quality improvements for end users
  - Improvements for translators
- The Future

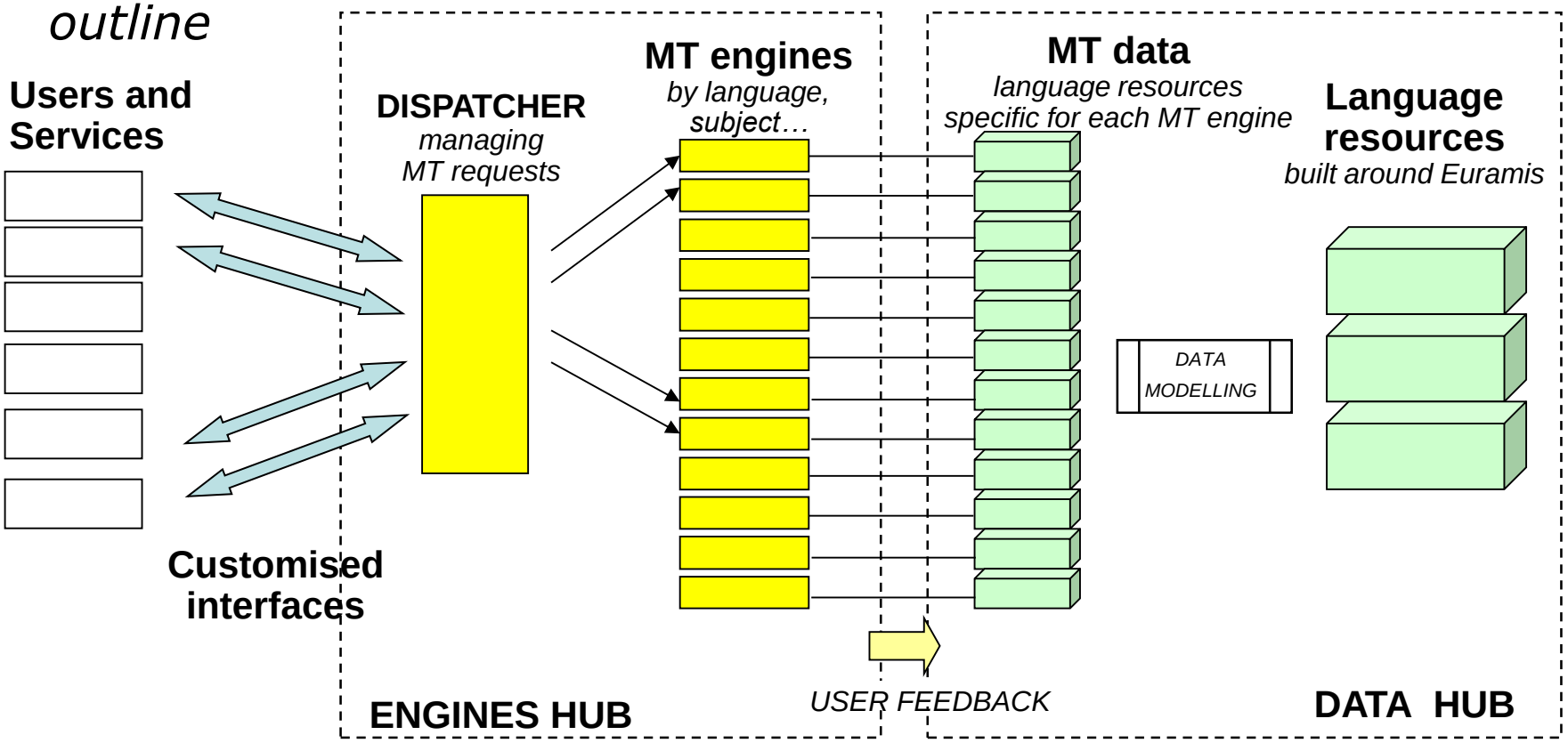
What are MT@EC and CEF.AT?

MT@EC:

- *Since **2010**, replacing an obsolescent rule-based solution (ECMT)*
- *Aimed at officials in EU and member states' administrations, both **translators** and **end users***
- *Covering all **24 EU languages** in all combinations (76 LPs directly)*
- *Based on open source Statistical MT technology (**Moses**), co-funded by EU Framework Programmes for research and innovation*
- *Developed by **DGT R3.2**, using co-funding by ISA programme*
- *Real-life trial with DGT translators since 2011, released in **June 2013***
- *Has been used to translate **tens of millions of pages***
  (up to 0.5M pages on a single day!)
- *Usage and feed-back from **translators** are one of the main sources of inspiration for **improvements** of MT **quality***

MT@EC project architecture

outline



MT action lines 3. Service 2. Engines 1. Data

What are MT@EC and CEF.AT?

CEF = Connecting Europe Facility

*Union financial assistance to **trans-European networks** in order to support projects of common interest in the sectors of transport, **telecommunications** and energy infrastructures and to exploit potential synergies between those sectors. Resources are to be made available under the multiannual financial framework for the years 2014-2020*

CEF.AT will:

- *build on the **existing** MT@EC service **but** not be limited to it*
- *put emphasis on **secure**, **quality**, **customisable** MT for pan-European online services (DSIs) **but** not be limited to them*
- *be a **multilingualism enabler, not only MT***

Platform is being built to serve

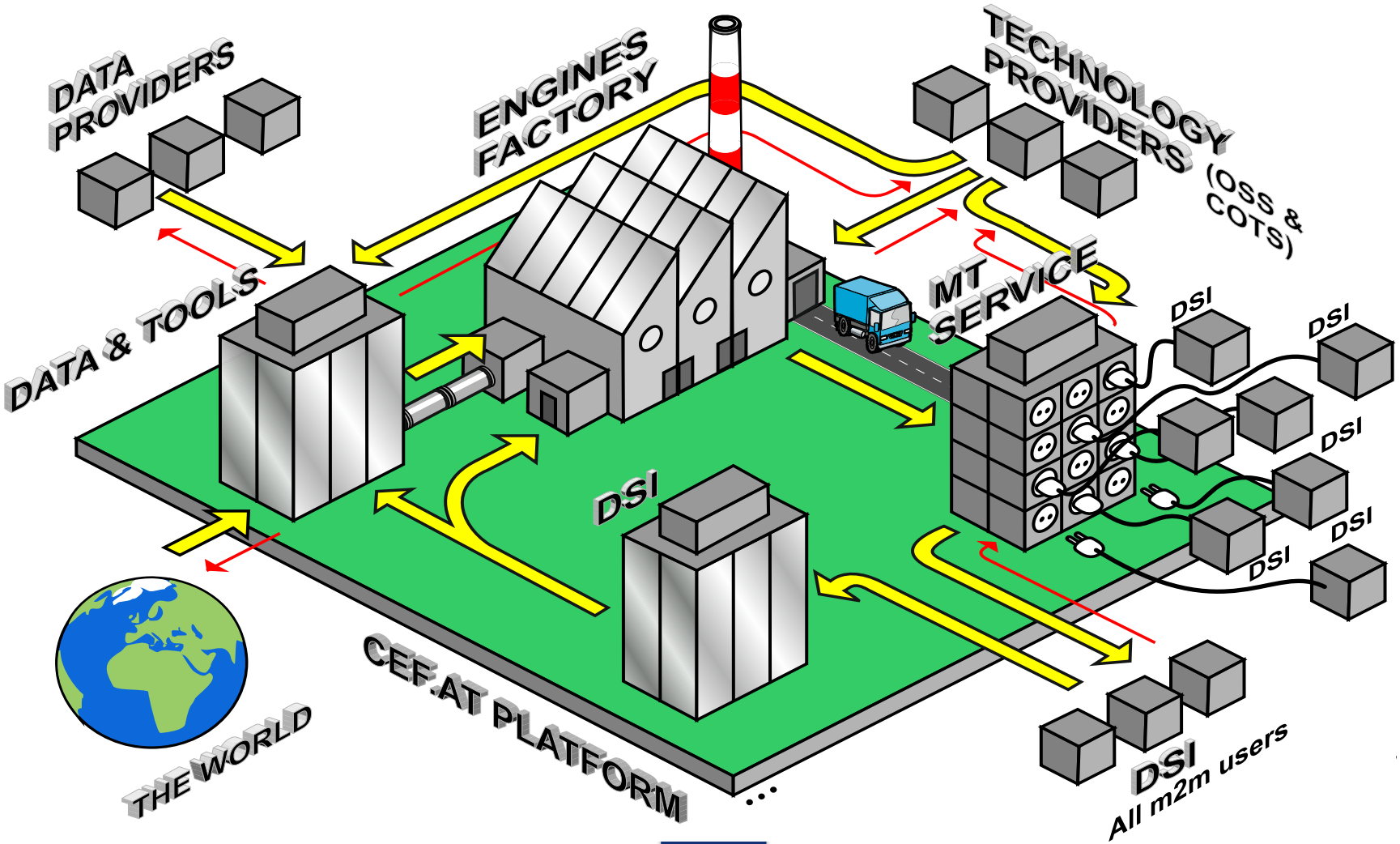
CEF Digital Service Infrastructures (DSI), other public online services, public bodies in the EU Member States, and European institutions/bodies

DSIs using CEF.AT

Initially we will serve 6 DSIs; the list may grow during the runtime of CEF

Service	Description
Europeana	The digital European Library, common, multilingual access point to digital resources of European heritage.
ODP	The pan-European **O**pen **D**ata **P**ortal for accessing open data infrastructures distributed over a EU and MS data repositories.
EESSI	The **E**lectronic **E**xchange of **S**ocial **S**ecurity **I**nformation, a platform between 32 countries (EU+EFTA).
ODR	The **O**nline **D**ispute **R**esolution platform for resolution of online contractual disputes between consumers and traders, linking all national Alternative Dispute Resolution (ADR) entities.
e-justice	A portal which is a single point of access to law, enabling EU judicial cooperation.
SaferInternet	Services to make Internet a trusted environment for children.

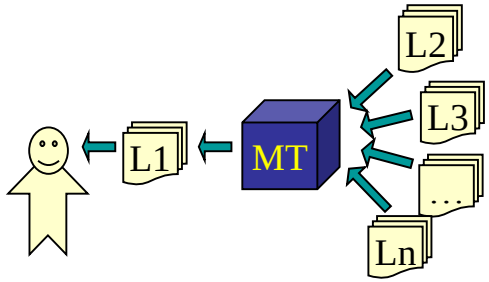
Building MT Engines for CEF



Main Usage Scenarios of MT

Requirements depend on the way MT is being used

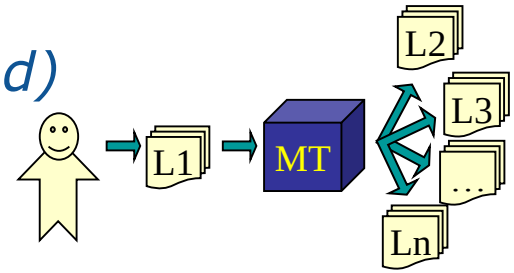
a) MT for end users (assimilation, inbound)



Robustness
Coverage, Scale

Practically unlimited demand; but free web-based services reduce incentive to improve technology

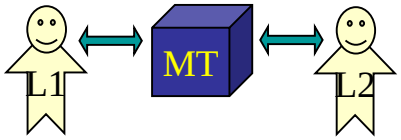
b) MT for translators (dissemination, outbound)



Textual quality, terminology, accuracy

Publishable quality can only be authored by humans; MT needs to be embedded into CAT Tools

c) MT for direct communication



Ill-formed input, recognition errors, specific style (chat), context dependence

MT as a module in larger information systems covering specific scenarios

Main Requirements

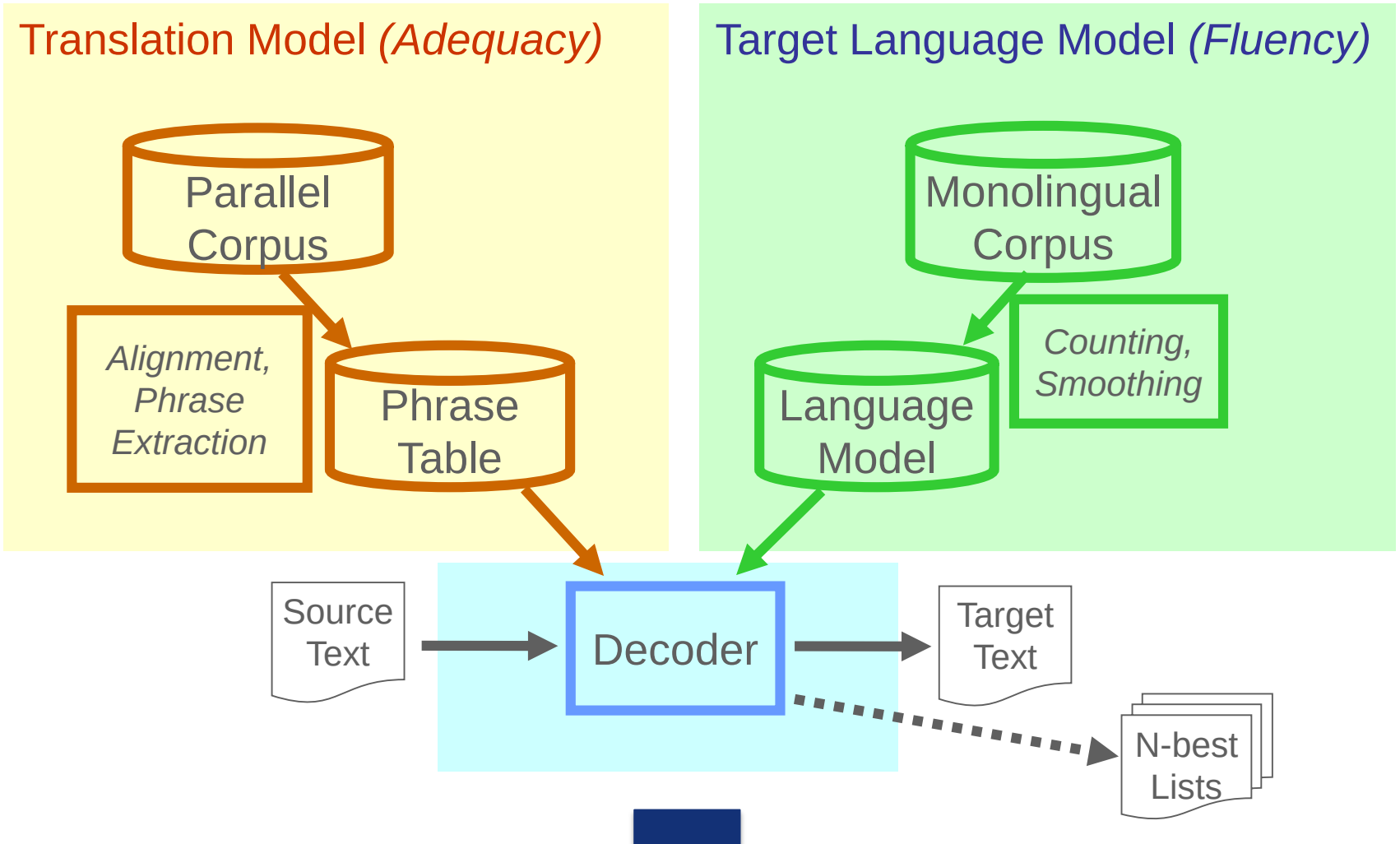
For end users (scenario a), EC and other administrations

- **Provide MT as a (simple and robust) service**
- **Optimise quality for understandability (gisting)**
- **Deal with many domains, document types, formats, …**
- **Scale to huge volumes**

For outbound translation (scenario b), DGT & similar institutions

- **Provide MT as a tool within a CAT workflow**
- **Develop new ways to incorporate feed-back from translators**
  - explicit feed-back on MT quality
  - implicit feed-back via TM
  - improvements requiring language-specific knowledge
  - towards linguistically informed/hybrid approaches
- **Optimise quality for post-editing**

Basic Architecture for Statistical MT



Word Alignment:

			CLASSIC SOUPS	Sm.	Lg.
清	燉	雞 湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞	飯	湯	58. Chicken Rice Soup	1.85	3.25
雞	麵	湯	59. Chicken Noodle Soup	1.85	3.25
廣	東	雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃	茄	蛋湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲	吞	湯	62. Regular Wonton Soup	1.10	2.10
酸	辣	湯	63. Hot & Sour Soup	1.10	2.10
蛋	花	湯	64. Egg Drop Soup	1.10	2.10
雲	蛋	湯	65. Egg Drop Wonton Mix	1.10	2.10
豆	腐	菜湯	66. Tofu Vegetable Soup	NA	3.50
雞	玉	米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹	肉	玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海	鮮	湯	69. Seafood Soup	NA	3.50

Word Alignment:

CLASSIC SOUPS

		Sm.	Lg.
清燉雞湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58. Chicken Rice Soup	1.85	3.25
雞麵湯	59. Chicken Noodle Soup	1.85	3.25
廣東雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃茄雲湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62. Regular Wonton Soup	1.10	2.10
酸辣湯	63. Hot & Sour Soup	1.10	2.10
雲花湯	64. Egg Drop Soup	1.10	2.10
雲雲湯	65. Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66. Tofu Vegetable Soup	NA	3.50
雞玉米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69. Seafood Soup	NA	3.50

Sample from a Phrase Table: EN → DE

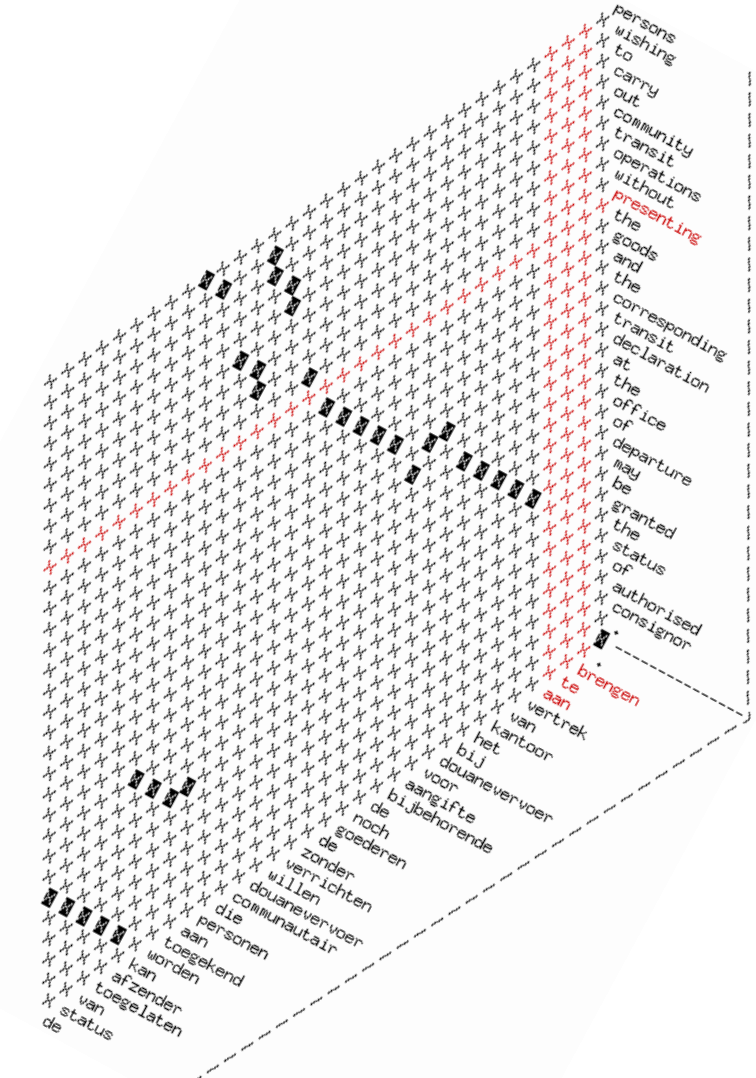
```
competition ||| wettbewerb ||| 0.792718 0.824898 0.471616 0.386947 2.718 ||| 0-0 ||| 84532 142086 67010
competition ||| wettbewerbs ||| 0.820595 0.871534 0.119012 0.0982444 2.718 ||| 0-0 ||| 20607 142086 16910
competition ||| den wettbewerb ||| 0.649377 0.824898 0.0612376 0.0100777 2.718 ||| 0-1 ||| 13399 142086 8701
competition ||| des wettbewerbs ||| 0.309026 0.871534 0.0220711 0.00127634 2.718 ||| 0-1 ||| 10148 142086 3136
competition ||| auswahlverfahren ||| 0.196723 0.150756 0.0212125 0.0211993 2.718 ||| 0-0 ||| 15321 142086 3014
competition ||| der wettbewerb ||| 0.578915 0.824898 0.0199809 0.0323991 2.718 ||| 0-1 ||| 4904 142086 2839
competition ||| auswahlverfahrens ||| 0.394891 0.346946 0.0166448 0.0161751 2.718 ||| 0-0 ||| 5989 142086 2365
competition ||| konkurrenz ||| 0.575426 0.544313 0.0133159 0.0130549 2.718 ||| 0-0 ||| 3288 142086 1892
competition ||| wettbewerbsrechtlichen ||| 0.528922 0.587997 0.00804442 0.009912 2.718 ||| 0-0 ||| 2161 142086 1143
competition ||| wettbewerbsbedingungen ||| 0.137052 0.149663 0.00745323 0.0205685 2.718 ||| 0-0 ||| 7727 142086 1059
competition ||| wettbewerbs- ||| 0.448318 0.580939 0.00515885 0.0047116 2.718 ||| 0-0 ||| 1635 142086 733
competition ||| wettbewerbspolitik ||| 0.156701 0.509962 0.00481399 0.0293835 2.718 ||| 0-0 ||| 4365 142086 684
competition ||| wettbewerb zu ||| 0.489235 0.824898 0.00463804 0.0107091 2.718 ||| 0-0 ||| 1347 142086 659
competition ||| wettbewerbsrechtliche ||| 0.314149 0.454091 0.00276593 0.0041319 2.718 ||| 0-0 ||| 1251 142086 393
competition ||| , den wettbewerb ||| 0.531609 0.824898 0.00260406 0.00132519 2.718 ||| 0-2 ||| 696 142086 370
competition ||| competition ||| 0.939394 0.787368 0.00239995 0.0021256 2.718 ||| 0-0 ||| 363 142086 341
competition ||| wettbewerbspolitik zuständige ||| 0.751724 0.509962 0.00230142 2.00395e-06 2.718 ||| 0-0 ||| 435 1420
competition ||| für wettbewerbspolitik ||| 0.603383 0.509962 0.0022592 0.000419673 2.718 ||| 0-1 ||| 532 142086 321
competition ||| konkurrieren ||| 0.115158 0.113537 0.00220289 0.0022166 2.718 ||| 0-0 ||| 2718 142086 313
competition ||| für wettbewerbspolitik zuständige ||| 0.71659 0.509962 0.00218882 2.86217e-08 2.718 ||| 0-1 ||| 434 14
competition ||| dem wettbewerb ||| 0.171412 0.824898 0.00207621 0.00260667 2.718 ||| 0-1 ||| 1721 142086 295
competition ||| wettbewerbsrechts ||| 0.10087 0.490932 0.00204102 0.0153851 2.718 ||| 0-0 ||| 2875 142086 290
competition ||| führen ||| 0.00311379 0.0028466 0.00190026 0.0025746 2.718 ||| 0-0 ||| 86711 142086 270
competition ||| wettbewerbsfähigkeit ||| 0.00331525 0.0036332 0.00187914 0.0015914 2.718 ||| 0-0 ||| 80537 142086 267
competition ||| wettbewerbsrecht ||| 0.0941485 0.483305 0.00176654 0.0211425 2.718 ||| 0-0 ||| 2666 142086 251
competition ||| wettbewerb , ||| 0.0736518 0.824898 0.00168208 0.0508825 2.718 ||| 0-0 ||| 3245 142086 239
competition ||| führt ||| 0.00342564 0.0027133 0.00150613 0.0017107 2.718 ||| 0-0 ||| 62470 142086 214
competition ||| wettbewerb herrscht ||| 0.291549 0.426895 0.00145686 0.000545402 2.718 ||| 0-0 0-1 ||| 710 142086 207
competition ||| wettbewerbsdruck ||| 0.0640314 0.05135 0.00137945 0.0017619 2.718 ||| 0-0 ||| 3061 142086 196
competition ||| wettbewerb in ||| 0.0863615 0.824898 0.00128795 0.0112427 2.718 ||| 0-0 ||| 2119 142086 183
...
7941 ENTRIES, covering 142086 occurrences!
```

Total size of this Phrase Table: 502 M entries

Sample from a Phrase Table (EN ➔ DE, II)

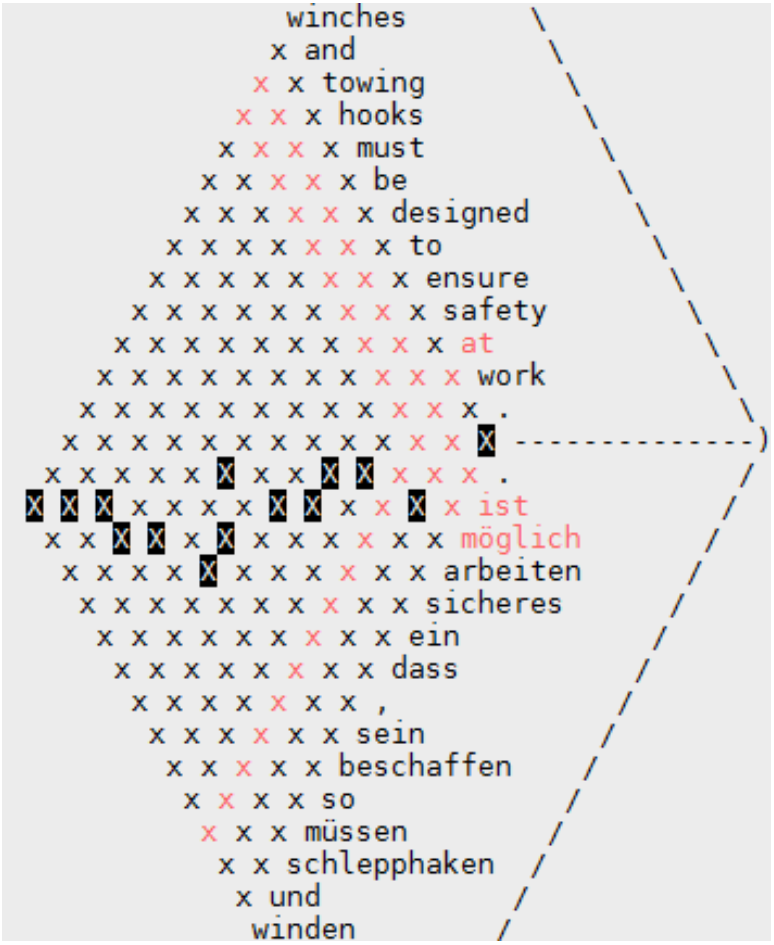
```
competition ||| , des bundeskartellamts ||| 0.0880911 0.0314685 3.71991e-06 8.74674e-08 2.718 ||| 0-2 ||| 6 142086 1
competition ||| , derweil der wettbewerb ||| 0.528547 0.824898 3.71991e-06 1.27812e-09 2.718 ||| 0-3 ||| 1 142086 1
competition ||| , der der konkurrenz ||| 0.528547 0.544313 3.71991e-06 1.20352e-05 2.718 ||| 0-3 ||| 1 142086 1
competition ||| , der den wettbewerbsregeln ||| 0.528547 0.443897 3.71991e-06 1.10122e-05 2.718 ||| 0-3 ||| 1 142086 1
competition ||| , den wettbewerb zu beschränken ||| 0.0101644 0.824898 3.71991e-06 1.34967e-09 2.718 ||| 0-2 ||| 52 142086 1
competition ||| , den wettbewerb in ||| 0.00997258 0.824898 3.71991e-06 3.85031e-05 2.718 ||| 0-2 ||| 53 142086 1
competition ||| , den wettbewerb in der ||| 0.0440456 0.824898 3.71991e-06 3.22387e-06 2.718 ||| 0-2 ||| 12 142086 1
competition ||| , den wettbewerb auf ||| 0.00960994 0.824898 3.71991e-06 1.55564e-05 2.718 ||| 0-2 ||| 55 142086 1
competition ||| , den wettbewerb auf dem ||| 0.0176182 0.824898 3.71991e-06 1.04796e-07 2.718 ||| 0-2 ||| 30 142086 1
competition ||| , dass wettbewerbsprobleme ||| 0.0528547 0.413662 3.71991e-06 3.63395e-06 2.718 ||| 0-2 ||| 10 142086 1
competition ||| , dass es ||| 7.32526e-06 3.4e-06 3.71991e-06 2.77654e-08 2.718 ||| 0-2 ||| 72154 142086 1
competition ||| , dass es zu einer unproduktiven konkurrenz ||| 0.528547 0.544313 3.71991e-06 0 2.718 ||| 0-6 ||| 1 142086 1
competition ||| , dass es sie ||| 0.00788876 3.4e-06 3.71991e-06 1.35692e-10 2.718 ||| 0-2 ||| 67 142086 1
competition ||| , dass einem wettbewerb ||| 0.528547 0.824898 3.71991e-06 7.98821e-07 2.718 ||| 0-3 ||| 1 142086 1
competition ||| , dass ein wettbewerb ||| 0.105709 0.824898 3.71991e-06 1.90358e-06 2.718 ||| 0-3 ||| 5 142086 1
competition ||| , dass die übernahme den wettbewerb ||| 0.528547 0.824898 3.71991e-06 3.30452e-11 2.718 ||| 0-5 ||| 1 142086 1
competition ||| , dass die früher ||| 0.0755067 0.0002227 3.71991e-06 1.1282e-09 2.718 ||| 0-3 ||| 7 142086 1
competition ||| , dass das konkurrieren ||| 0.528547 0.113537 3.71991e-06 1.43608e-08 2.718 ||| 0-3 ||| 1 142086 1
competition ||| , dass aufgrund ||| 0.00030676 8.2e-06 3.71991e-06 1.67179e-08 2.718 ||| 0-2 ||| 1723 142086 1
competition ||| , das heisst die teilnahme ||| 0.528547 0.0019528 3.71991e-06 2.3896e-14 2.718 ||| 0-4 ||| 1 142086 1
competition ||| , das dem wettbewerb ||| 0.132137 0.824898 3.71991e-06 2.27147e-06 2.718 ||| 0-3 ||| 4 142086 1
competition ||| , damit der wettbewerb ||| 0.031091 0.824898 3.71991e-06 2.46123e-06 2.718 ||| 0-3 ||| 17 142086 1
competition ||| , behindere den wettbewerb ||| 0.528547 0.824898 3.71991e-06 7.95114e-10 2.718 ||| 0-3 ||| 1 142086 1
competition ||| , aus denen der wettbewerb ||| 0.528547 0.824898 3.71991e-06 3.06829e-08 2.718 ||| 0-4 ||| 1 142086 1
competition ||| , also um zu vermeiden ||| 0.264273 0.0006296 3.71991e-06 3.39252e-13 2.718 ||| 0-4 ||| 2 142086 1
competition ||| ) wettbewerbs ||| 0.105709 0.871534 3.71991e-06 0.000688575 2.718 ||| 0-1 ||| 5 142086 1
competition ||| ) die konkurrenz ||| 0.176182 0.544313 3.71991e-06 6.17479e-06 2.718 ||| 0-2 ||| 3 142086 1
competition ||| ) den wettbewerb ||| 0.0587274 0.824898 3.71991e-06 7.06326e-05 2.718 ||| 0-2 ||| 9 142086 1
competition ||| ) darstellen ||| 0.00352364 0.0007677 3.71991e-06 1.47395e-06 2.718 ||| 0-1 ||| 150 142086 1
competition ||| ) besteht ||| 0.000597228 0.0001879 3.71991e-06 1.5931e-06 2.718 ||| 0-1 ||| 885 142086 1
competition ||| ( vgl. ||| 1.62126e-05 5.93e-05 3.71991e-06 9.22169e-08 2.718 ||| 0-1 ||| 32601 142086 1
competition ||| ( pro auswahlverfahren ||| 0.176182 0.150756 3.71991e-06 3.74981e-09 2.718 ||| 0-2 ||| 3 142086 1
competition ||| ( pro auswahlverfahren ) ||| 0.528547 0.150756 3.71991e-06 2.62817e-11 2.718 ||| 0-2 ||| 1 142086 1
competition ||| ( european competition ||| 0.0440456 0.787368 3.71991e-06 3.61082e-10 2.718 ||| 0-2 ||| 12 142086 1
competition ||| ( derzeit immer intensiver werdenden ) wettbewerbs ||| 0.528547 0.871534 3.71991e-06 4.44457e-25 2.718 ||| 0-6 |||
competition ||| ( d. h. wettbewerb ||| 0.528547 0.824898 3.71991e-06 4.18807e-07 2.718 ||| 0-2 ||| 1 142086 1
competition ||| ( aeuv ) darstellen ||| 0.105709 0.0007677 3.71991e-06 1.51821e-13 2.718 ||| 0-3 ||| 5 142086 1
```

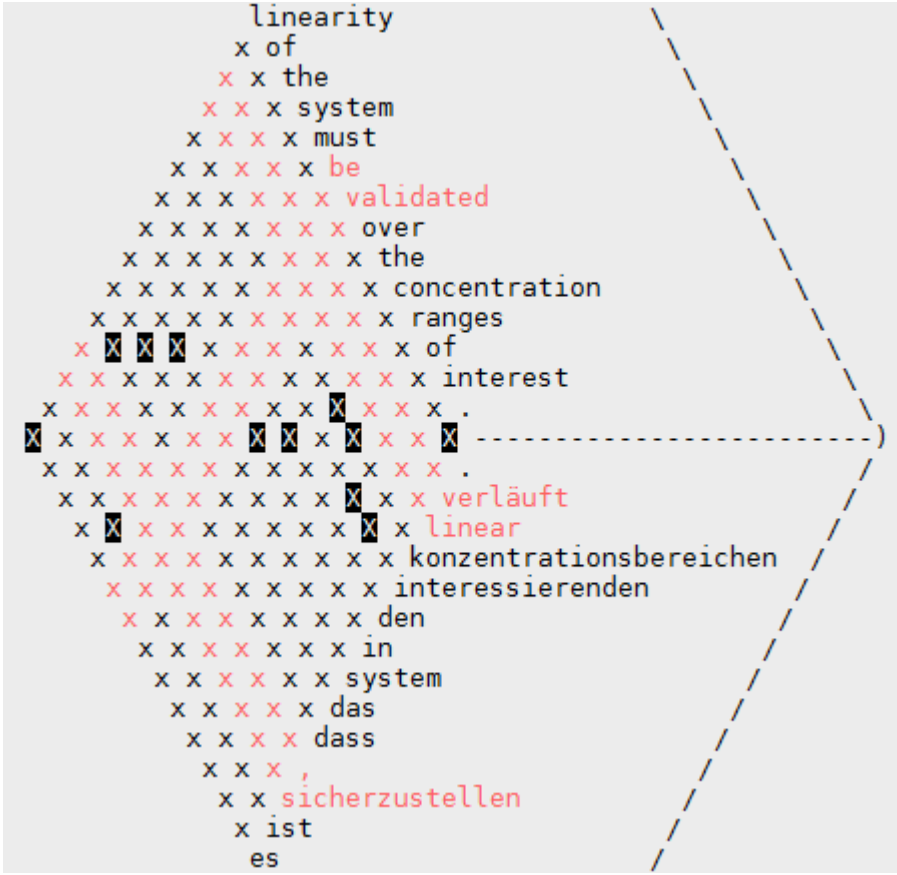
Where do wrong phrase table entries come from?



without presenting ||| zonder
without presenting the ||| zonder de
without presenting the goods ||| zonder de goederen
without presenting the goods and ||| zonder de goederen noch
without presenting the goods and the ||| zonder de goederen noch de
without presenting the goods and the corresponding
||| zonder de goederen noch de bijbehorende
presenting the ||| de
presenting the goods ||| de goederen
presenting the goods and ||| de goederen noch
presenting the goods and the ||| de goederen noch de
presenting the goods and the corresponding
||| de goederen noch de bijbehorende
the office of departure ||| het kantoor van vertrek aan te brengen
office of departure ||| kantoor van vertrek aan te brengen
of departure ||| van vertrek aan te brengen
departure ||| vertrek aan te brengen
. ||| aan te brengen .
. ||| te brengen .
. ||| brengen .

More examples of misalignment





Observations about typical errors

Typical errors depend mainly on target language (TL)

- Morphologically simple TL: Statistical models work reasonably well
- Strongly inflected TL: Word endings are often wrong
- Differences in order between SL and TL → more alignment errors → spurious deletions & insertion

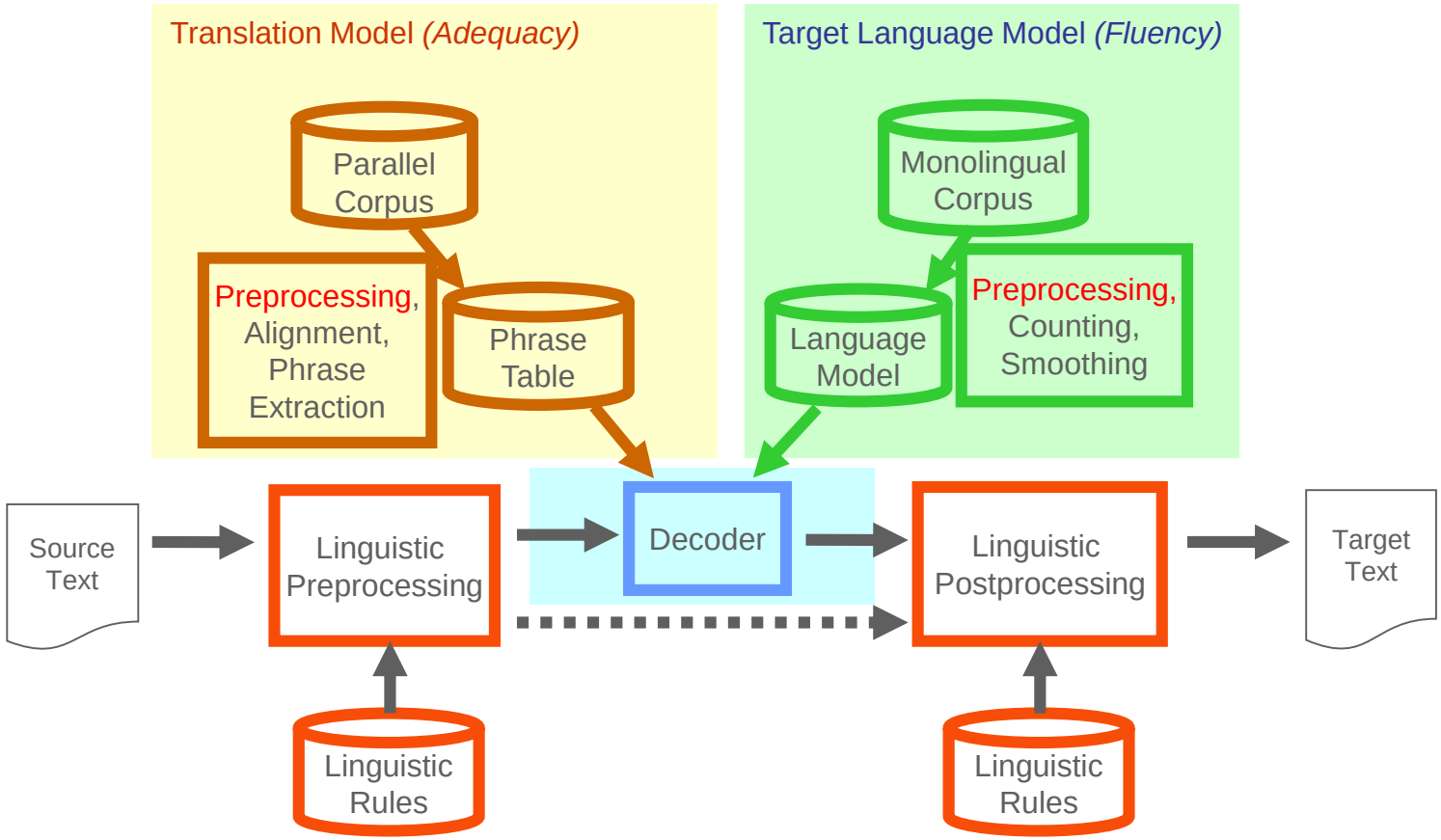
Some frequent errors can be fixed with simple means

- Certain types of expressions can be treated with rules
- Normalising punctuation helps a lot

Errors caused by different word order can be reduced

- Re-ordering before alignment reduces alignment errors
- Effect on final translation quality positive, but no breakthrough
- Increases complexity of software infrastructure → not yet in use

Enhanced Architecture for Hybrid (Rule-Based +Statistical) MT



Ways to improve translation quality

- Collect feed-back on frequent errors, refine rule-based modules
  ongoing work based on "Language Weeks" within DGT
- Integrate linguistic tools in pre- and post-processing/PT pruning
  part-of-speech models, morphology, parsing, reordering, …
- Distinguish relevance of training data
  domain awareness, recency, data quality indicators, …
- More data
  typically improves coverage, but may hurt disambiguation and domain-awareness if not done well
- Different types of data
  Lexicons, Terminologies, Ontologies, …

How to better serve the needs of end users

- Build models optimised for different use cases (**domain adaptation**), e.g. specific MT engines for DSIs that have sufficient training data
- Improve **scalability**
  - Better capacity and response times via cloud computing
  - Offer choice between speed and accuracy
- Better **coverage** of general-purpose vocabulary
- Better **robustness** when dealing with low-quality input
- **Linguistic improvements** will also help end users

How to better serve the needs of translators

- Ongoing: **Implement improvements** identified during language weeks (see next slides)
- Work on **domain adaptation** should also help translators (Euramis covers many different domains)
- **Learn from** stream of **corrections** (implicit feed-back) done by translators using the system
- Better integration into **CAT environment** (e.g. Auto-Suggest functionality in SDL Studio)

More on Language Weeks

- LDs provided lists of observations ("issues")
- MT team analysed and classified them into four classes:
  - **A: simple bug that can be easily fixed**
  - **B: doable in the short term but requires considerable effort**
  - **C: solution may be feasible but it may not achieve any improvement or will require much more effort**
  - **D: the existing technology is not mature enough for a fix to be applicable at a reasonable cost**
- An optional **+** was used to mark issues where we need additional input from LD

MT issues collected during Language Weeks

- Some 700 issues were analysed and classified
- Distribution over classes looks as follows:

	Fixed	A	B	C	D
X	20	3	67	130	376
X+		1	43	64	
	20	4	110	194	376
	2.8%	0.6%	15.6%	27.6%	53.4%

- Class A/B and some class C issues for 18 languages were so far recorded in JIRA bug tracker ➔ 57 tickets, often relevant for many or all languages
- 22 improvements were implemented before training of 10th generation engines started in November 2015

Recent improvements in 10th generation

- spaces in Lithuanian dates
- page abbreviation in Polish
- straight quotes in Danish
- reworked normalisation of dashes/hyphens for all languages, especially when in between numbers
- replaced Swedish word "skall" by "ska"
- all languages: placeholder contents are now uppercased at the beginning of a segment
- "fx" abbreviation in Danish
- non-breaking spaces in Lithuanian abbreviations
- Maltese numbers and articles
- Lithuanian alternative endings
- inflected placeholders (Czech, Slovak, Croatian)
- Danish corrupted characters
- Slovak term for "amendments"
- French number filter
- segment-level casing (all languages)
- link filtering (all languages)
- Euro -> EUR (all languages)
- no space before numbers/placeholders ("M1") (all languages)
- English date placeholder extended to "31st of December" etc.
- non-breaking spaces in German dates
- "decision" -> "Beschluss" in German
- no comma after opening bracket (all languages)

Towards better integration into CAT

Baseline:

TMX or XLIFF from MT@EC used by CAT tool

PRO: *Simple to implement (real-life trial since 2011, official version since June 2013), may fit different CAT tools*

CON: *Post-editing tedious, alternative translation choices considered during MT decoding not accessible to user*

Improvement (in work):

*Feed chart of translation **options** into CAT tool*

PRO: *may be more helpful than single MT result*

CON: *Requires the development of a plug-in for Studio and suitable formats for generalised MT results*

Next Steps

- **We have been given the chance to embed MT@EC into a much bigger initiative "Connecting Europe Facility"**
  - New types of big users with specific needs for domain adaptation
  - This allowed us to enlarge the size of the MT Engines team (going from 4 to 8 computational linguists)
  - Access to Cloud computing infrastructure

- **Follow-up work on "Language Weeks"**
  - First changes of the MT system based on LWs were included in the 10th and 11th generations of MT engines, more are in work
  - Many more improvements will come during the year, some will require us to collaborate with translators on the fine details

- **Embedding in translators' work flow**
  - CAT tool integration
  - Collect and analyse changes made in post editing

The Future

- The MT research community at large is working towards a better **integration of linguistic knowledge** sources
- Language resources collected in member states via **ELRC** will be integrated into our MT solution
- Other big trends are: ***Big Data*** (corpora crawled from the Web) and ***Deep Learning*** (artificial neural networks to obtain better statistical models)
- We observe these developments closely and try out relevant new techniques on our data if possible
- As soon as some new method becomes mature enough to be included, we will work on doing so

Questions?

andreas.eisele@ec.europa.eu
DGT-MT@ec.europa.eu