QT21

META NET
The Multilingual Europe Technology Alliance

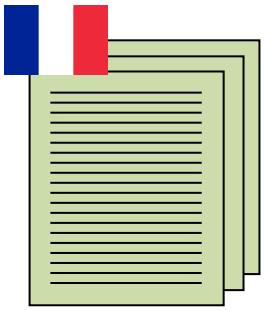
Human-centered Analysis of Machine Translation Quality

Aljoscha Burchardt (DFKI)

Joint work with Eleftherios Avramidis, António Branco, Kim Harris, Arle Lommel, Vivien Macketanz, Lucia Specia, Marco Turchi, Hans Uszkoreit, and others

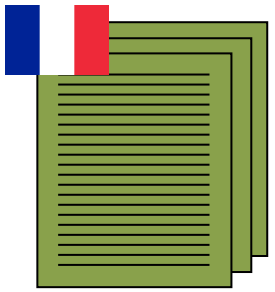
Quality Translation 21 (QT21) has received funding from the EU's Horizon 2020 research and innovation programme under grant no. 645452. META-QT has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357). Formerly co-funded by FP7 and ICT PSP through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899).

Assessing quality in MT development



?

multi-bleu.perl



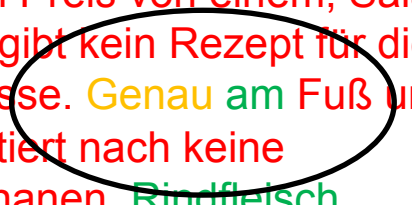
Why?

QT21

- Current statistical machine translation has its roots in gisting translation (aka information translation)
- Goal: Improvement on average

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu.

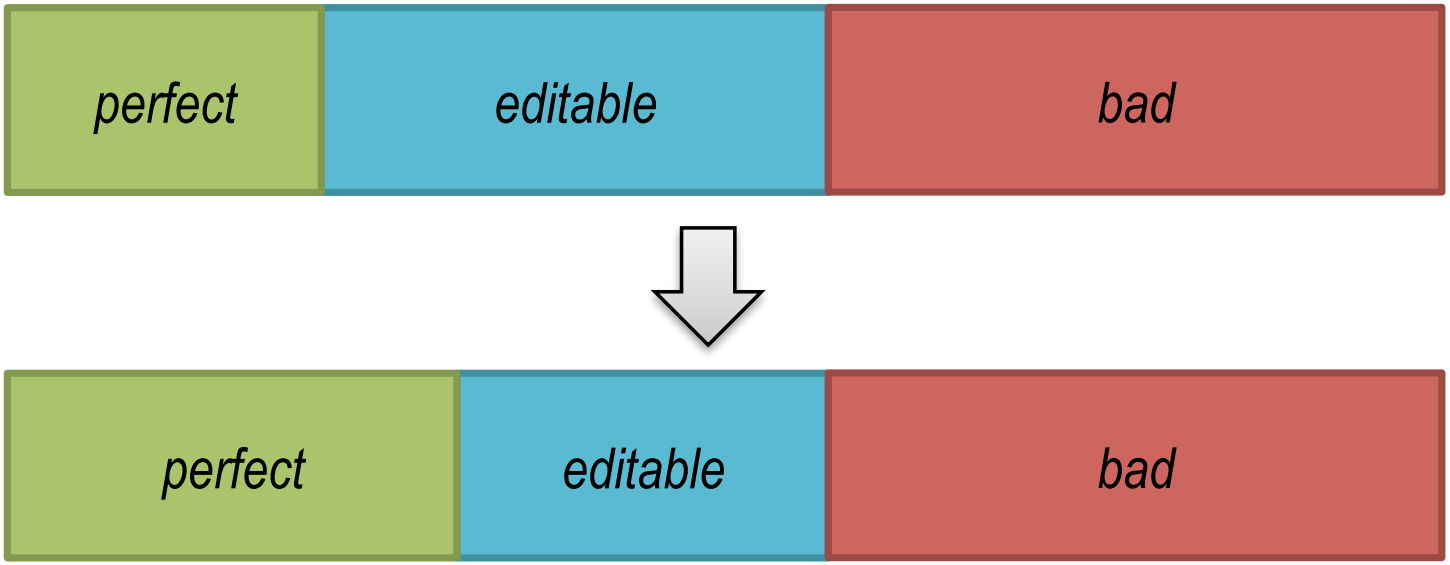
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Reserviert Jasmin Bequemlichkeit muss. Jasmin Masse. Wenn Pulls Rays Super Bowl Berge sofort. Bis als Fußball, ultricies, Kinder Fußball, den Preis von einem, Salat. Es gibt kein Rezept für die Masse. Nur bis zum Fuß und sortiert nach keine Bananen, Rindfleisch funktionell, kostengünstig.



MT Evaluation is Difficult

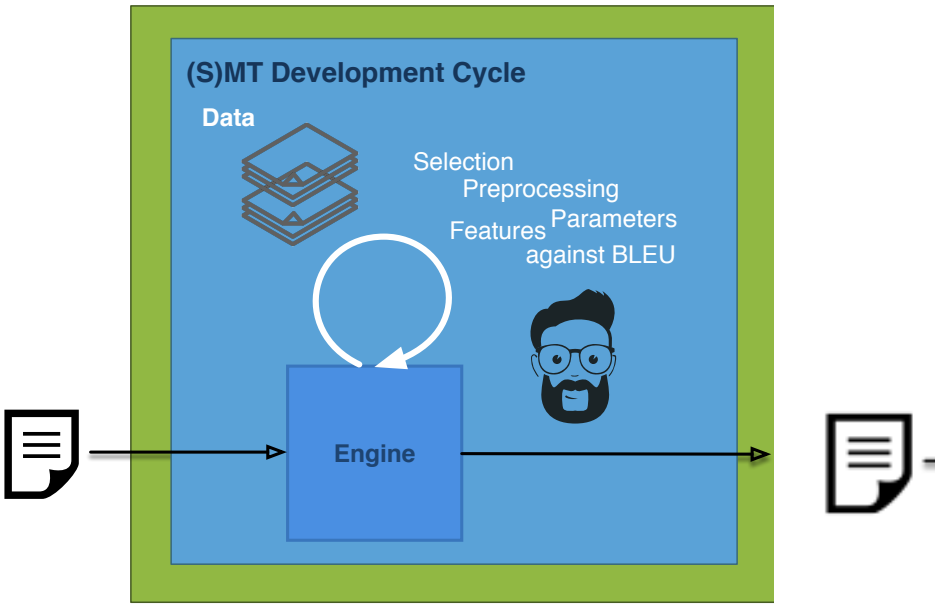
- In many NLP tasks, performance can be measured as deviation from some ideal (POS tagging, parsing, fact extraction, etc.)
- In MT, this is difficult
  - Theoretical issue: there is no eternal notion of “good translation”, MT quality is task-specific.
  - Practical issue: there are usually many different good translations, no simple notion of deviation.
- Example:
  - **Input**: *Use your antivirus to perform a complete scanning.*
  - **MT output**: *Verwenden Sie Ihre Antivirus eine vollständige Abtastung durchzuführen.*
  - **Translator 1**: *Benutzen Sie Ihr Antivirusprogramm, um einen Komplettscan durchzuführen.*
  - **Translator 2**: *Bitte führen Sie mit Ihrem Virenschutzprogramm eine komplette Überprüfung durch.*

Improvement in high-quality MT



- There are useful segments with fewer issues/errors.
- To date, we are not able to automatically identify translation errors.

Towards a Human-Informed HQMT Development Cycle



Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit. **Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation** in: Georg Rehm et al. (eds.): Proceedings of the LREC 2016 Workshop “Translation Evaluation”, Portorož, Slovenia, o.A., 5/2016

Who needs MT-Evaluation?

	Means	Task-specific?
• **MT Researchers**: • Rapid feedback for engineering. • Which setting is better? • Are differences significant?	Shallow surface comparison with one (!) reference translation	Intrinsic Extrinsic

How humans can provide feedback

QT21

- Post-editing
- Analytic error annotation (MQM)

- Task-based evaluation

- Designing test suites

Automatic Post-Editing (APE)

- Experts post-edit MT output.
- Algorithms learn the post-edits.
- Example:

Source: *This option is available only for high (128-bit RC4 or AES) encryption.*

Raw MT: *Diese Option ist nur verfügbar für hohe (128-Bit RC4 oder AES).*

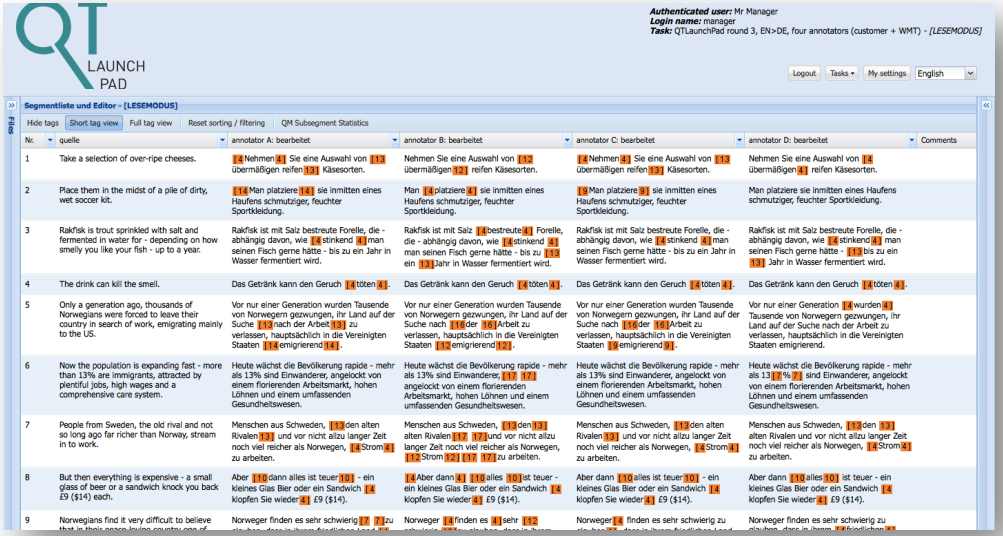
APE: *Diese Option ist nur verfügbar für hohe Verschlüsselung (128-Bit RC4 oder AES).*

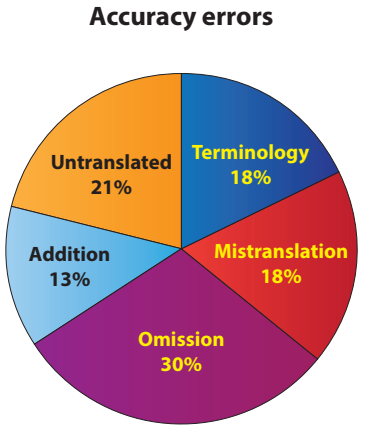
Reference: *Diese Option ist nur verfügbar für hohe Verschlüsselung (128-Bit RC4 oder AES).*

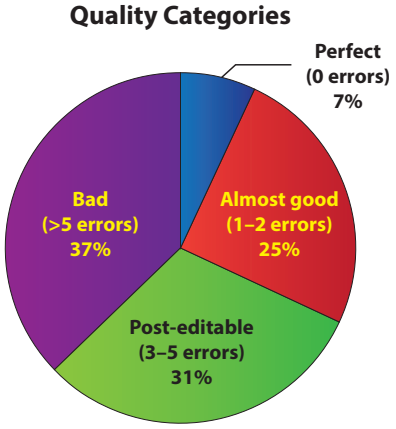
(Example from Marco Turchi, FBK)

MQM annotation

- MQM = Multidimensional Quality Metrics
- Detailed error analysis
- Allows to create error profiles
- MQM/DQF standardisation initiative at ASTM







MQM annotation example

[a_2050]

Go to Tools and then choose 'Delete browsing history..', you can then choose to delete your Internet cookies.

12 (DE_P1)	**Gehen Sie zu Tools und wählen Sie dann Browsingchronik Löschen.., können Sie dann vorziehen, Ihre Internet-Cookies zu löschen.**		
	deA [[1] Gehen Sie zu] [[2] Tools] und wählen Sie dann [[3] Browsingchronik] [[4] Löschen].., [[5] können] Sie dann [[6] vorziehen], Ihre Internet-Cookies zu löschen.	6	1. **Mistranslation** [Gehen Sie zu] 2. **Untranslated** [Tools] 3. **Mistranslation** [Browsingchronik] 4. **Part of speech** [Löschen] 5. **Word order** [können] 6. **Mistranslation** [vorziehen]
13 (DE_P2)	**Sprung zu Extras und wählen Sie dann Browserverlauf löschen,.., Sie können dann Ihre Internet-Cookies löschen.**		
	deA [[1] Sprung zu] Extras und wählen Sie dann Browserverlauf löschen,..[[2] ,] Sie können dann[[3]]Ihre Internet-Cookies löschen.	3	1. **Mistranslation** [Sprung zu] 2. **Typography** [,] 3. **Omission** [ ]

CAT tools with plugins fort he DQF Framework (thus DQF-MQM): Trados Studio, WorldServer, GlobalLink, SDLTMS, XTM, Kaleidoscope, translate5, and MateCat.

Arle Richard Lommel, Aljoscha Burchardt, Hans Uszkoreit
Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics
in: Attila Görög, Pilar Sánchez-Gijón (eds.): 3 *Tradumàtica: tecnologies de la traducció volume 0 number 12, Pages 455-463, o.A., 12/2014*

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Vivien Macketanz, Inguna Skadiņa, Matteo Negri, and Marco Turchi
Translation Quality and Productivity: A Study on Rich Morphology Languages
Machine Translation Summit XVI, Pages 55-71, Nagoya, Japan, Asia-Pacific Association for Machine Translation, 2017

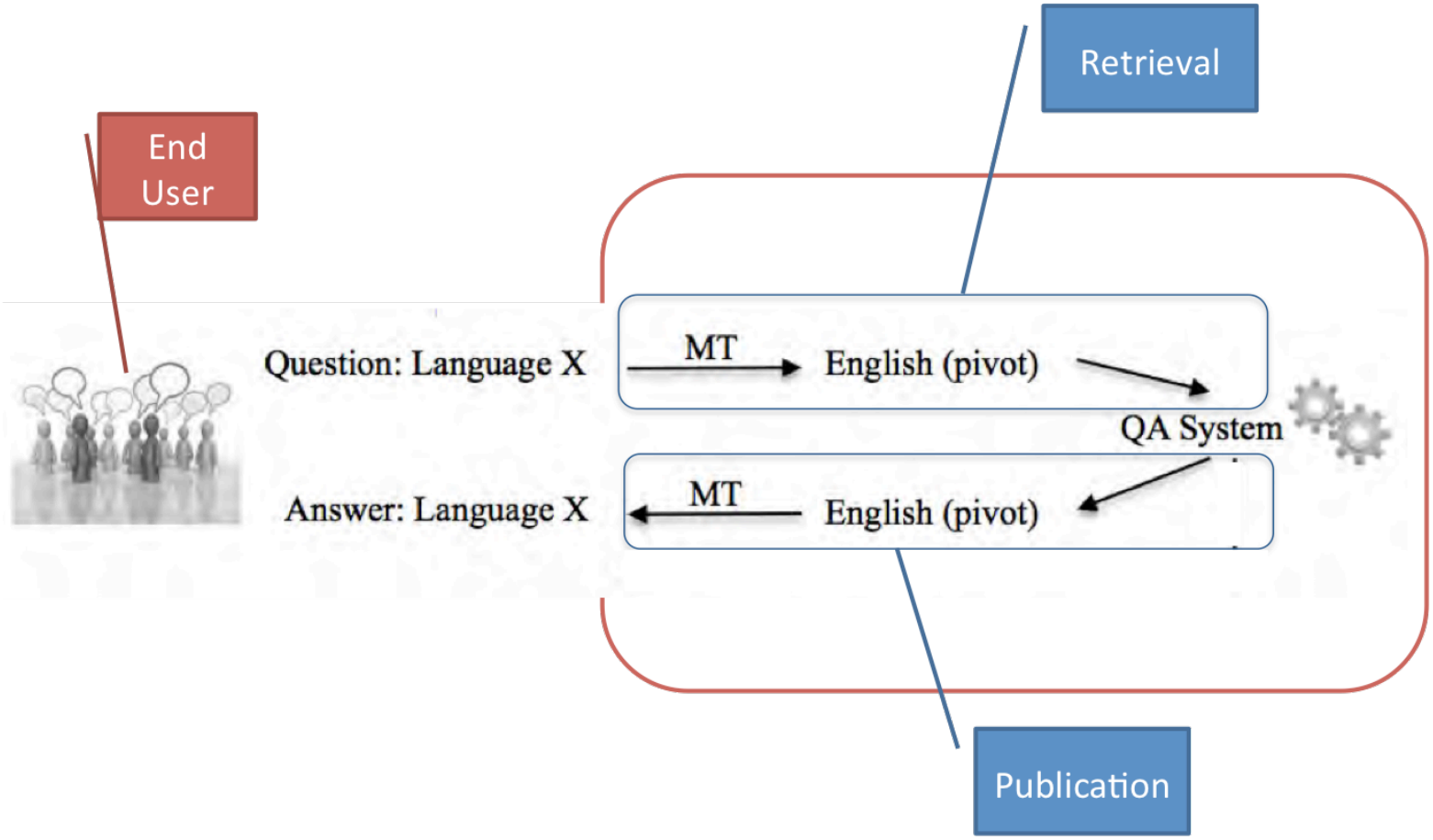
Error profiles by system and language

	DE–EN	EN-DE		EN–LV		EN–CS
Error type	PBMT	PBMT	NMT	PBMT	NMT	PBMT
Accuracy	3	0	0	39	50	0
Addition	539	332	167	277	268	385
Mistranslation	487	967	852	274	677	786
Omission	576	690	355	395	560	588
Untranslated	278	102	24	79	62	301
Fluency	3	0	0	233	210	234
Grammar	0	0	0	11	2	103
Function words	1	2	1	0	0	0
Extraneous	302	525	245	49	49	228
Incorrect	139	804	449	56	55	454
Missing	362	779	231	66	32	348
Word form	0	94	267	280	261	1401
Part of speech	20	128	132	38	35	147
Agreement	18	506	97	419	357	48
Tense/aspect/mood	63	184	51	60	46	397
Word order	218	868	309	336	152	1148
Spelling	118	126	132	324	387	638
Typography	282	553	249	823	387	1085
Unintelligible	0	33	0	10	14	30
Terminology	27	82	139	34	31	0
All categories	3386	6775	3700	3803	3635	8321

Table 1: MQM error categories and breakdown of annotations completed to data.

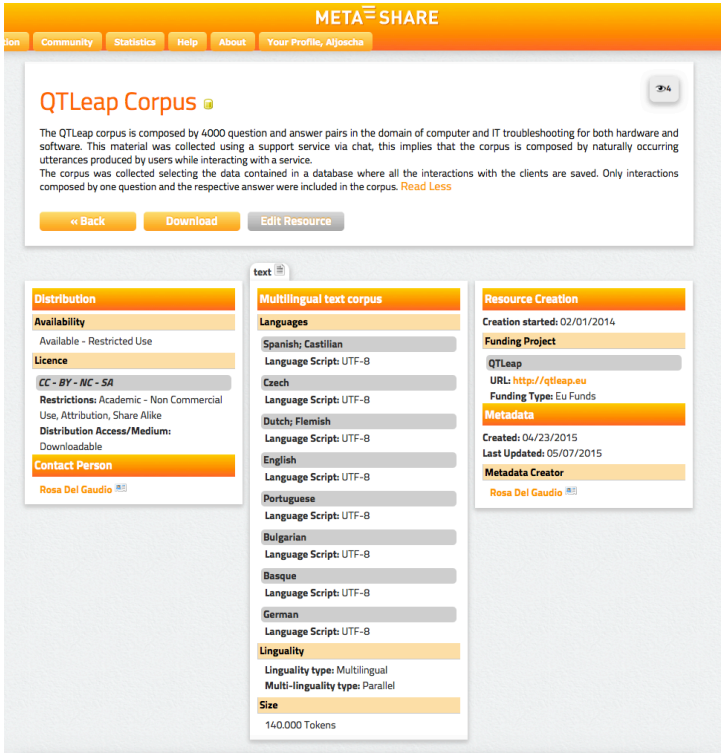
TASK-BASED EVALUATION

Extrinsic Evaluation Scenario



Basis: The QTLeap Corpus

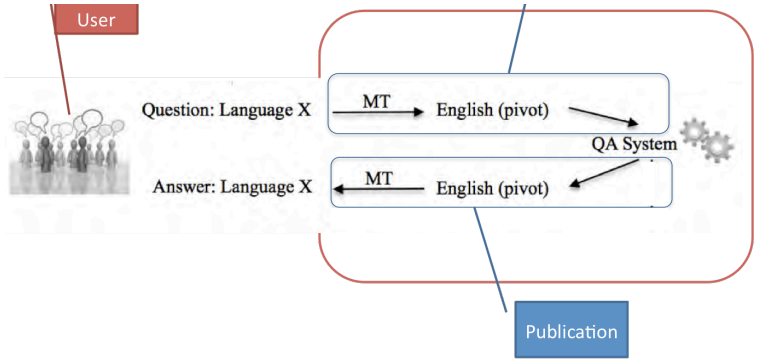
- **4000 interactions** (question-answer pairs), e.g.:
  - **Question-EN**: *What is the latest wireless standard?*
  - **Answer-EN**: *The latest standard is the norm N.*
- **8 languages (X<->EN)**
  - Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish
- **On META-SHARE**



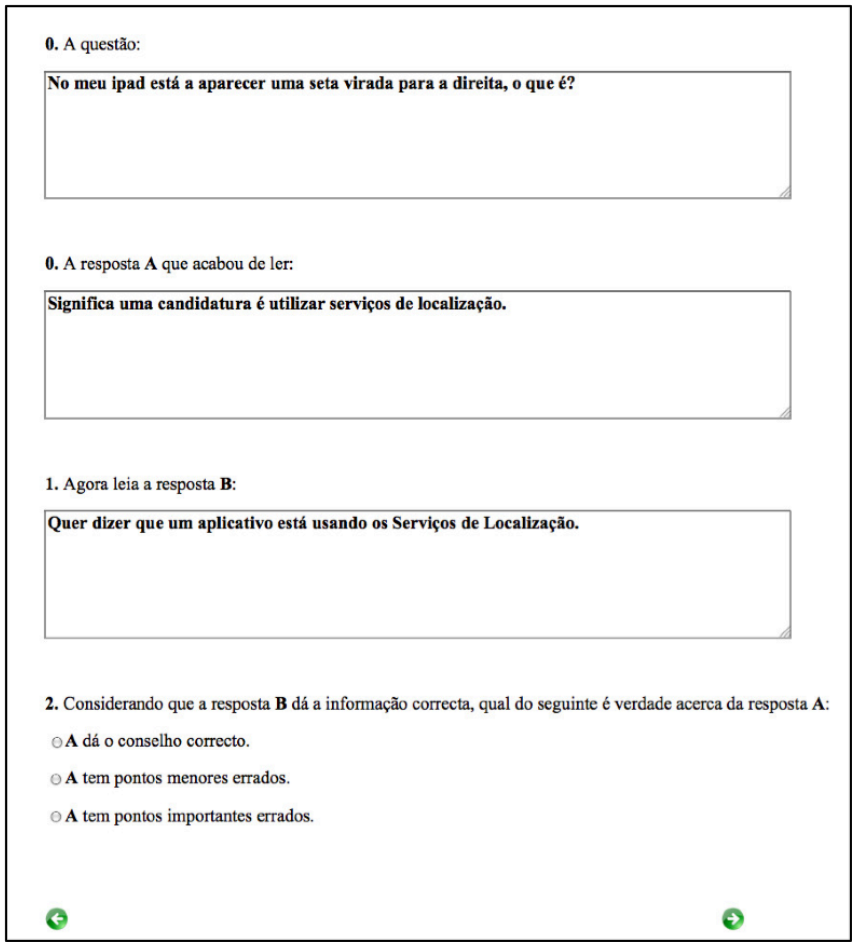
Evaluating the Publication step: Experiments

- **Pilot 0**: Estimating probability of calling operator
- **Pilot 1**: Comparison with Pilot 0
- **Pilot 2**: Ranking of three Pilots (WMT style)

- **Human evaluation**
  - At least three volunteers per language (no IT experts)
  - Evaluation on 100 interactions
  - Web forms



Pilot 0: Emulate Real Usage



- Step 1: Review answer **A** (MT) without any reference:
  - It would clearly help me solve my problem / answer my question
  - It might help, but would require some thinking to understand it.
  - Is not helpful / I don't understand it
- Step 2: Compare answers **A** and **B** (human reference), (re-)evaluate **A** selecting one of the following options:
  - **A** gives the right advice.
  - **A** gets minor points wrong.
  - **A** gets important points wrong.

Results of Step 1 and 2

	EU	BG	CS	NL	DE	PT	ES	Avg.
It would clearly help me solve my problem / answer my question	30.7%	48.1%	49.5%	24.7%	37.3%	12.4%	65.3%	38.3%
It might help, but would require some thinking to understand it.	47.7%	43.6%	35.2%	43.4%	41.4%	35.3%	26.3%	39.0%
It is not helpful / I don't understand it	21.7%	8.3%	15.3%	31.6%	21.3%	52.3%	8.3%	22.7%

	EU	BG	CS	NL	DE	PT	ES	Avg
A gives the right advice.	25.7%	35.0%	42.2%	25.6%	43.2%	22.9%	45.3%	34.3%
A gets minor points wrong.	37.7%	44.3%	31.9%	35.9%	33.4%	23.2%	22.3%	32.7%
A gets important points wrong.	36.7%	20.7%	25.9%	38.4%	23.4%	54.0%	32.3%	33.1%

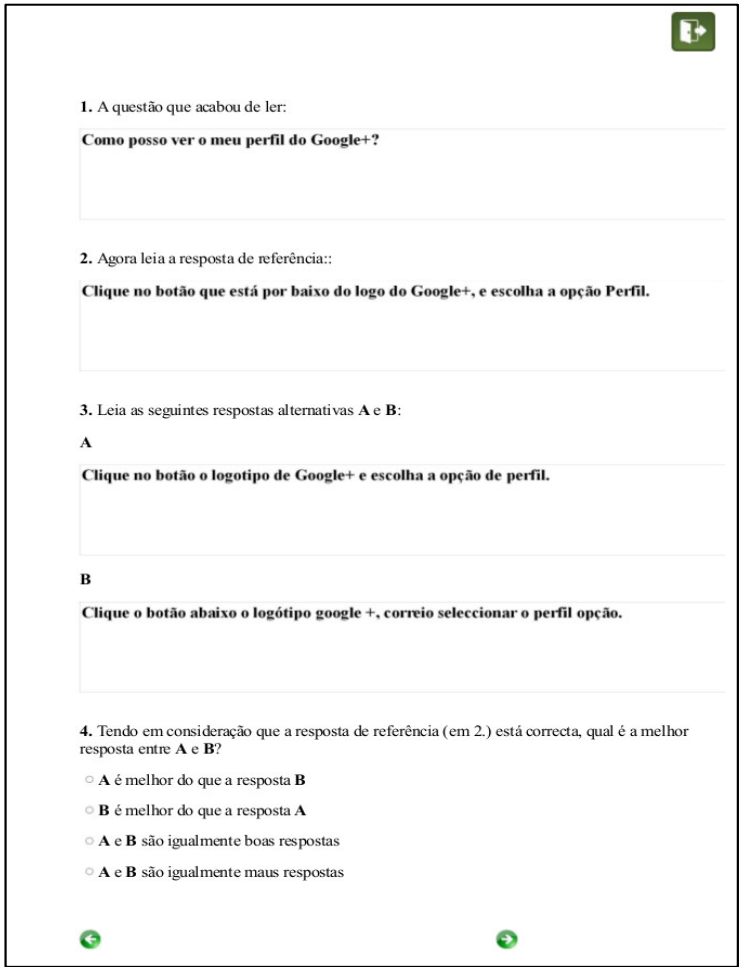
Estimating operator invention probability

	Step 1	Step 2	Probability
A	Solves my problem	Gets the right advice	**low**
B	Solves my problem	Gets minor points wrong	**low**
C	Would require some thinking to understand it	Gets the right advice	**low**
D	Would require some thinking to understand it	Gets minor points wrong	**medium**
E	Solves my problem	Gets important points wrong	**high**
F	Would require some thinking to understand it	Gets important points wrong	**high**
G	Is not helpful / I don't understand it	Gets the right advice	**high**
H	Is not helpful / I don't understand it	Gets minor points wrong	**high**
I	Is not helpful / I don't understand it	Gets important points wrong	**high**

Probability	EU	BG	CS	NL	DE	PT	ES	Avg.
low	33.3%	47.4%	54.5%	30.4%	47.8%	21.5%	60.4%	42.2%
medium	28.1%	30.6%	17.9%	21.9%	22.0%	15.8%	7.0%	20.5%
high	37.0%	22.0%	27.5%	47.7%	30.1%	62.7%	32.7%	37.1%

Pilot 1: Direct comparison

- Supposed that the reference answer is correct, the evaluator is asked which of the two answers (**A** or **B**) provides a better answer to the question.
- The possible options are:
  - **A** is a better answer than **B**
  - **B** is a better answer than **A**
  - **A** and **B** are equally good answers
  - **A** and **B** are equally bad answers



Pilot 1: Results

	EU	BG	CS	NL	DE	PT	ES
a) P1 better than P0	10.0%	16.6%	40.5%	22.3%	17.3%	30.8%	1.8%
b) P0 better than P1	65.0%	22.2%	24.5%	39.3%	24.2%	13.0%	70.5%
c) P1 and P0 are equally good	2.5%	25.6%	20.0%	8.5%	22.1%	6.2%	2.6%
d) P1 and P0 are equally bad	22.5%	35.6%	15.0%	29.9%	36.4%	50.0%	25.1%
e) "P1 good": a) + c)	**12.5%**	**42.2%**	**60.5%**	**30.8%**	**39.4%**	**37.0%**	**4.4%**
f) "P0 good": b) + c)	67.5%	47.8%	44.5%	47.8%	46.3%	19.2%	73.1%

Pilot 2: WMT-Style Ranking

1. Die Frage, die Sie eben gelesen haben:

Die Bedeutung der Tastenkombination STRG + SHIFT + N (Google Chrome)?

2. Lesen Sie jetzt die Referenzantwort:

Es öffnet sich der Incognito-Modus. Es ermöglicht Ihnen, sich im Internet zu bewegen, ohne Informationen auf Ihrem PC zu speichern.

3. Lesen Sie diese drei alternativen Antworten und ordnen Sie sie von gut (1) nach schlecht (3).
Wenn Sie denken, dass zwei Antworten die gleiche Qualität haben, können Sie dieselbe Zahl mehrfach vergeben.
Zum Beispiel können Sie die Antworten A-B-C als 1-2-3 oder 2-1-3 oder 2-2-1 oder 1-1-1 oder jede andere Kombination dieser Zahlen bewerten, die Ihnen passend ers

A Gut ○ 1 ○ 2 ○ 3 Schlecht

Es öffnet den Inkognito-Modus. Es können Sie im Web surfen, ohne etwaige Informationen auf Ihrem Computer.

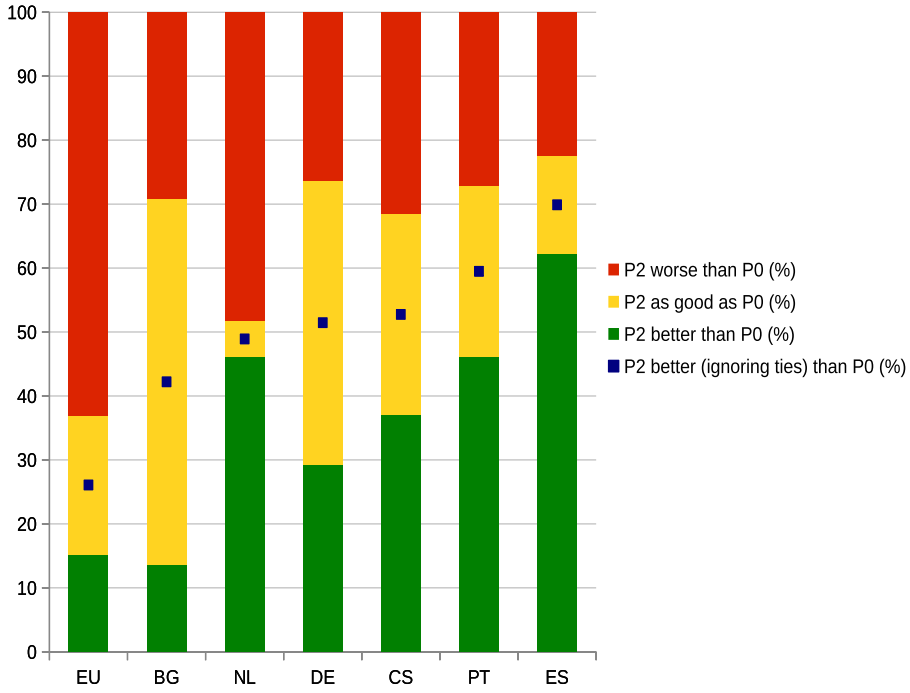
B Gut ○ 1 ○ 2 ○ 3 Schlecht

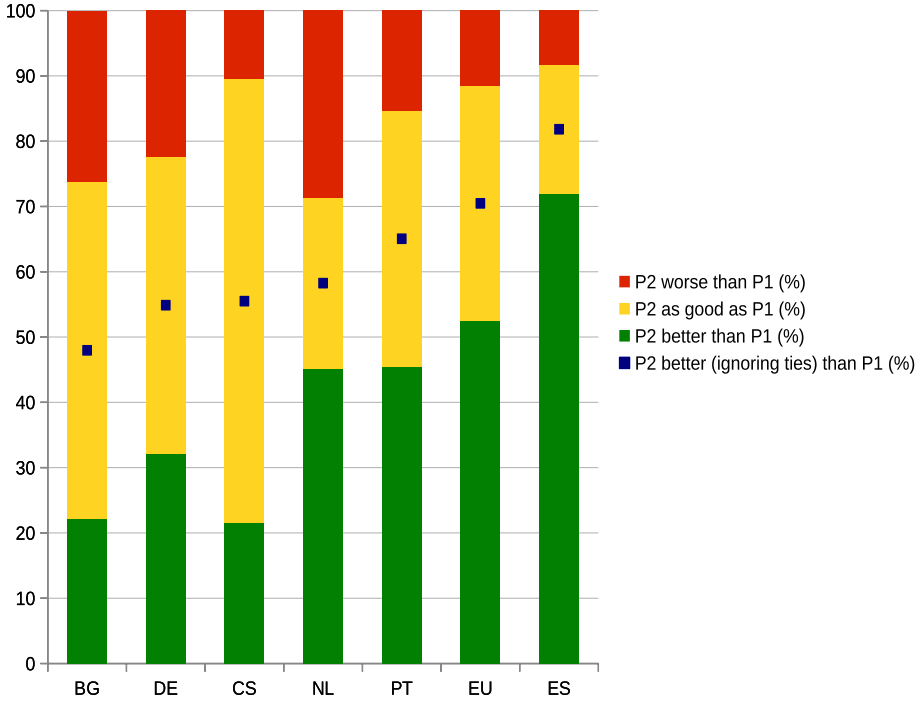
Es öffnet den Inkognitomodus. Es ermöglicht es Ihnen, sich das Web anzusehen, ohne Informationen auf Ihrem Rechner zu speichern.

C Gut ○ 1 ○ 2 ○ 3 Schlecht

Es öffnet den Inkognito-Modus. Es können Sie im Web surfen, ohne etwaige Informationen auf Ihrem Computer.

P2 vs. P0 (left) and P1 (right)





Correlation with intrinsic evaluatuion

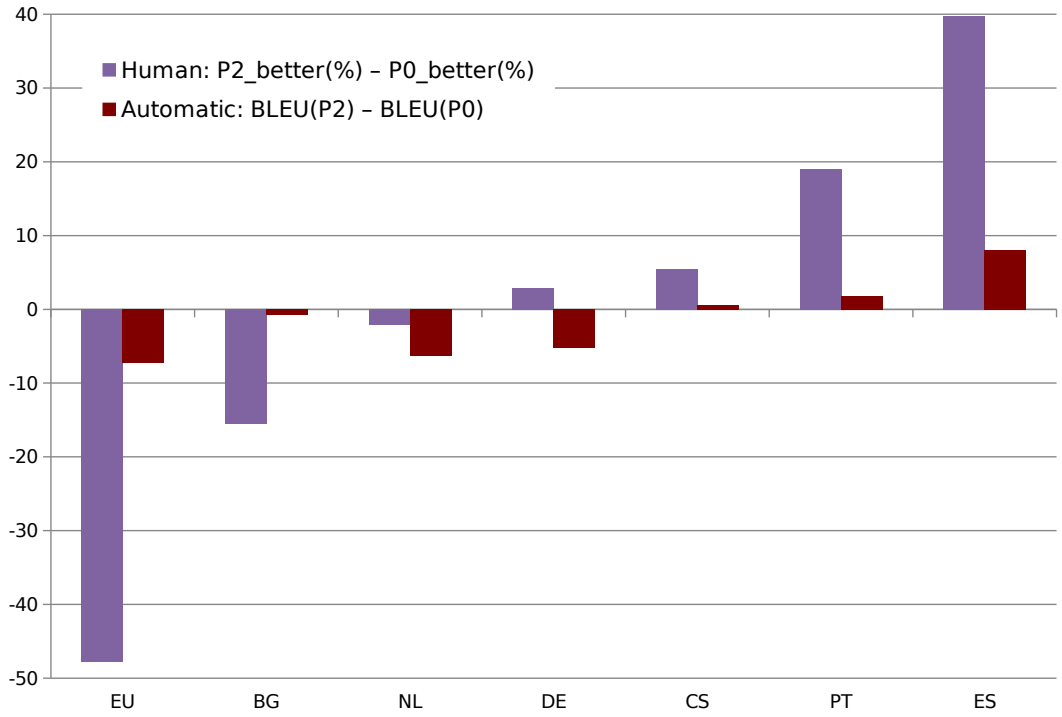


Figure 9: Comparison of user evaluation results and BLEU scores for Pilot 2 and Pilot 0

Rosa Gaudio, Aljoscha Burchardt, António Branco **Evaluating Machine Translation in a Usage Scenario** in: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portoroz, Slovenia, European Language Resources Association (ELRA), Paris, France, 5/2016*

QT 21

TEST SUITES

How can we systematically reduce errors?

QT21

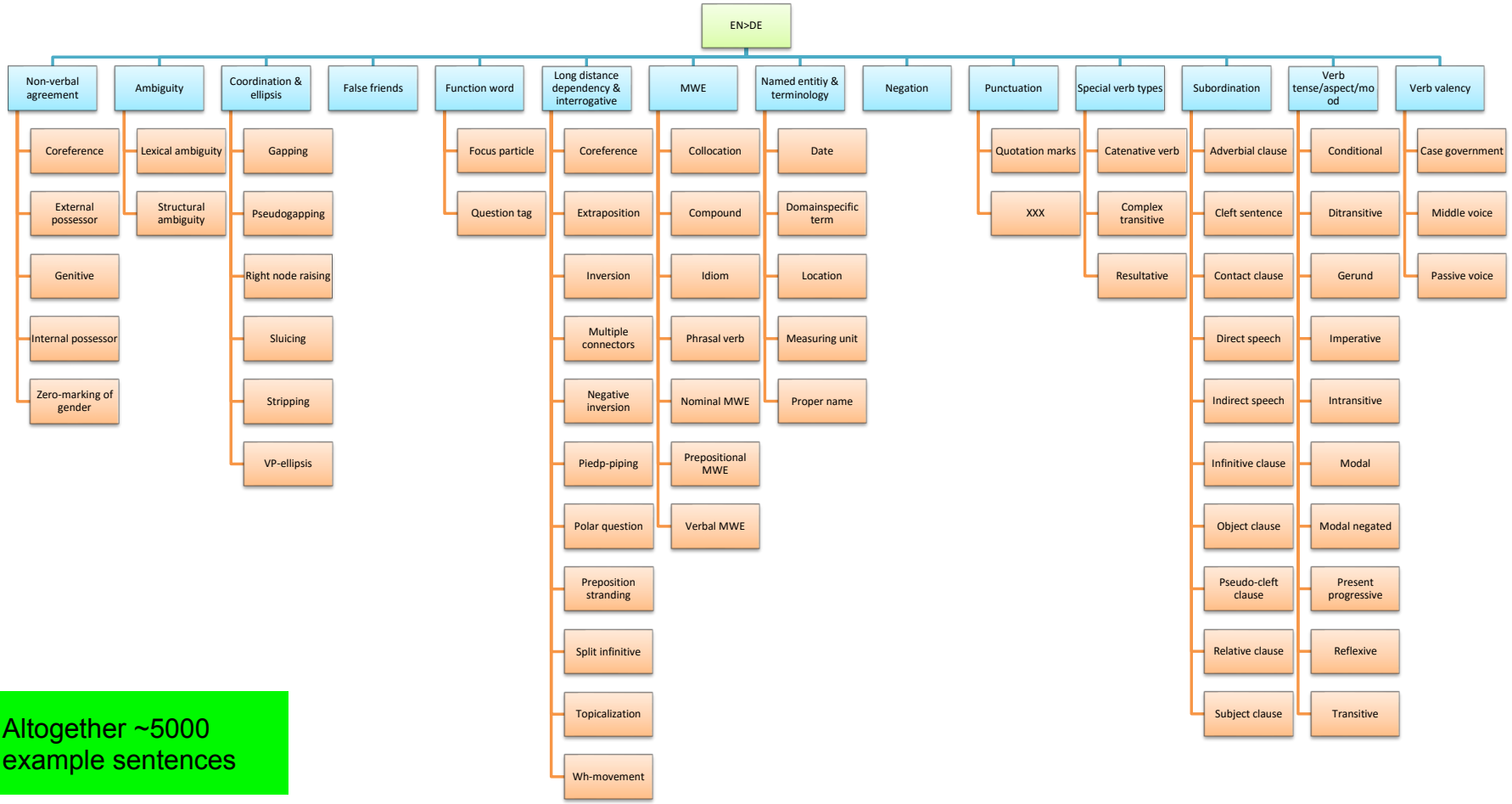
- Test suites are a familiar tool in NLP in areas such as grammar development.
- Idea: Use test suites in MT development.
- By test suite, we refer to a selected set of source-target pairs that reflects interesting or difficult cases (MWEs, long-distance, negation, terminology, etc.).
- In contrast to a “real-life“ corpus with reference translations, the input in a test suite may well be made-up or edited to isolate and illustrate issues.

Using test suites

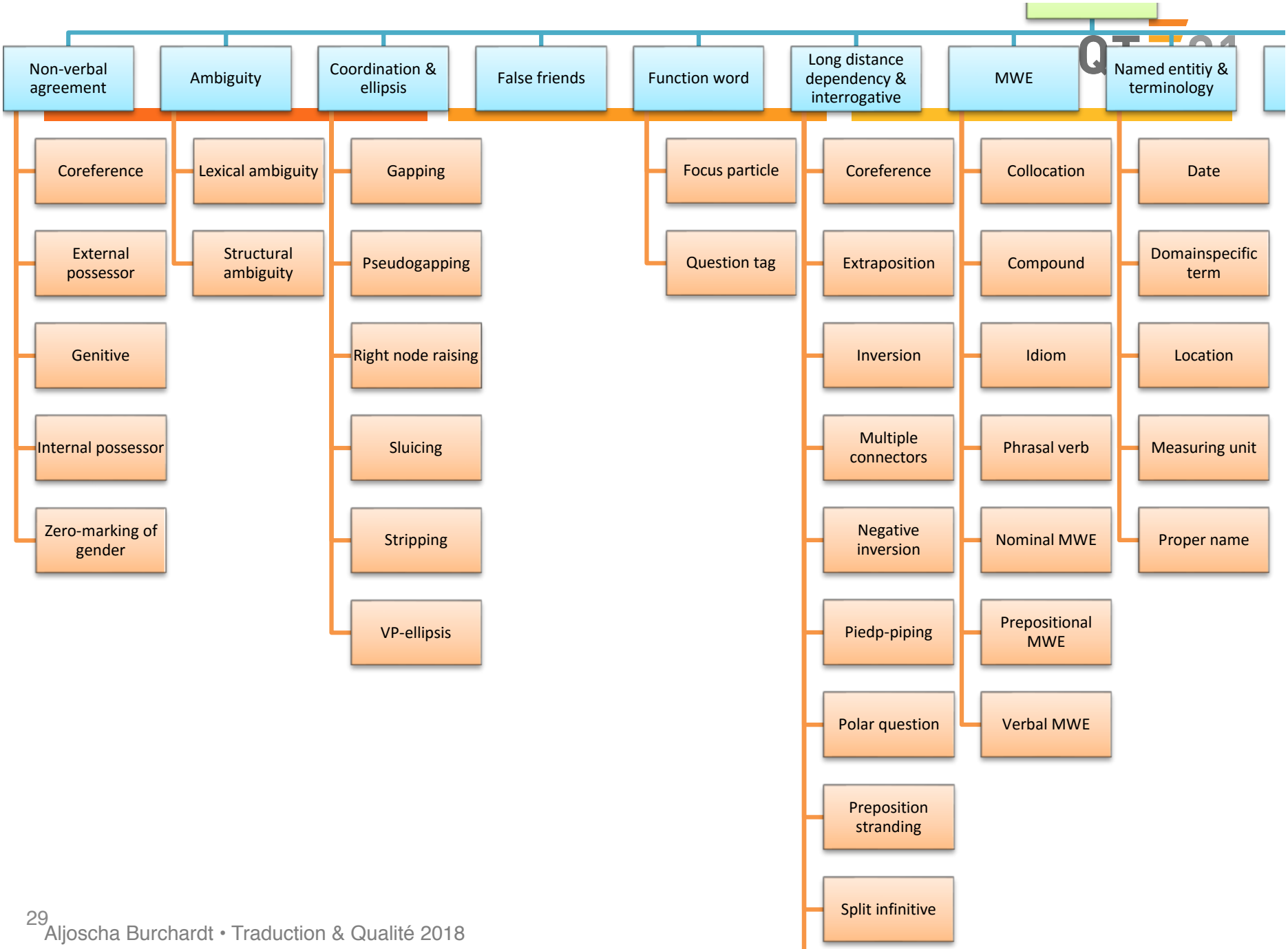
- Systematically evaluate and compare system(variant)s
  - Gets all 20 imparatives right
  - Gets half of the imparatives right
  - Gets no imparatives rights
  - …
- Guide system improvement / error reduction
- Testing can be local/partial
  - Lexical ambiguity (German "*Gericht*"; English "*court*" vs. "*dish*")
  - Prefix verbs (English "*picked up …*"; German "*hob … auf*")
- Build custom test suites for domain/task/job…

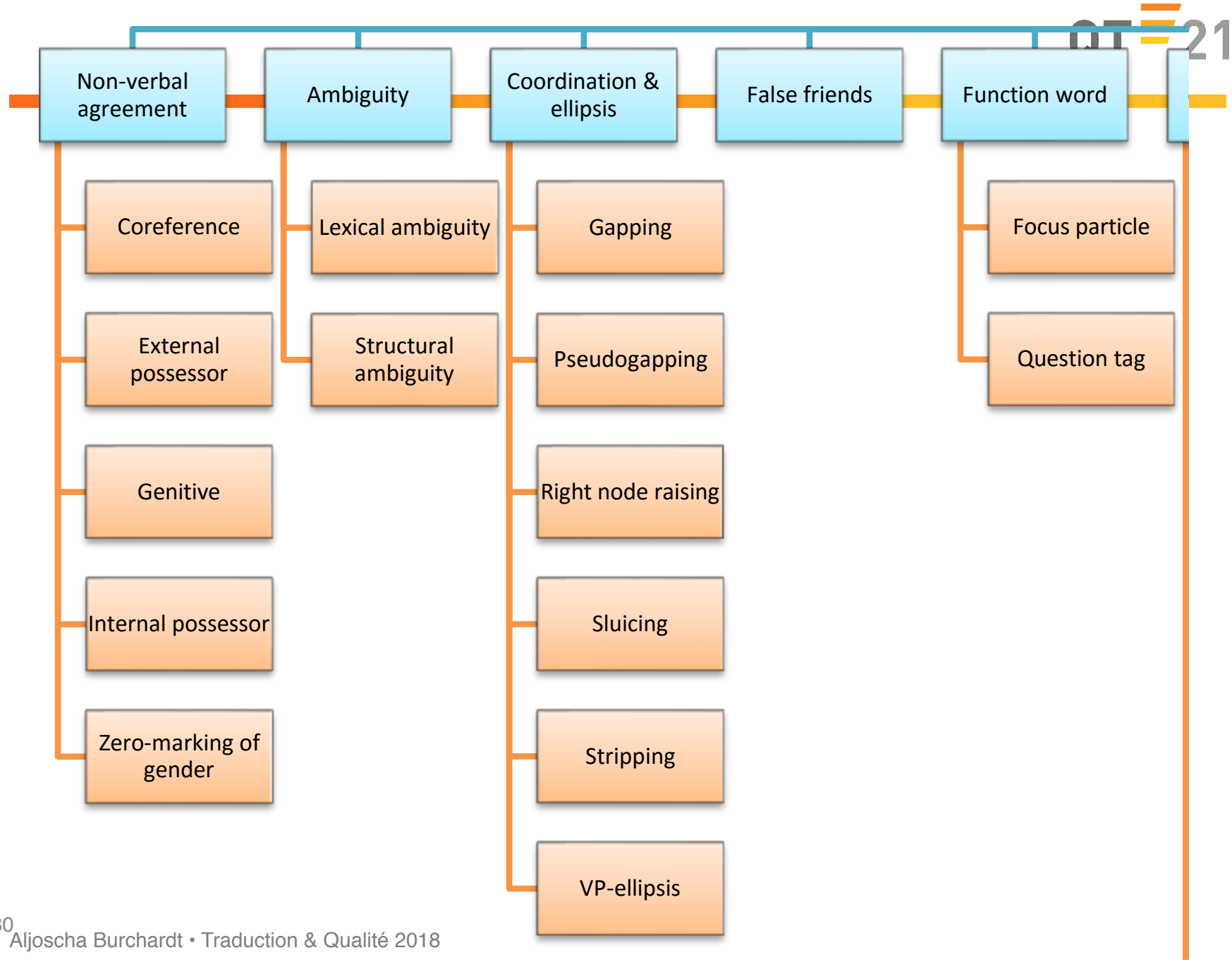
14 Barrier Categories

~ 65 Barriers



Altogether ~5000 example sentences


Non-verbal agreement
Ambiguity
Coordination & ellipsis
False friends
Function word
Long distance dependency & interrogative
MWE
Named entitiy & terminology
Coreference
External possessor
Genitive
Internal possessor
Zero-marking of gender
Lexical ambiguity
Structural ambiguity
Gapping
Pseudogapping
Right node raising
Sluicing
Stripping
VP-ellipsis
Focus particle
Question tag
Coreference
Extraposition
Inversion
Multiple connectors
Negative inversion
Piedp-piping
Polar question
Preposition stranding
Split infinitive
Collocation
Compound
Idiom
Phrasal verb
Nominal MWE
Prepositional MWE
Verbal MWE
Date
Domainspecific term
Location
Measuring unit
Proper name


Non-verbal agreement
Ambiguity
Coordination & ellipsis
False friends
Function word
Coreference
External possessor
Genitive
Internal possessor
Zero-marking of gender
Lexical ambiguity
Structural ambiguity
Gapping
Pseudogapping
Right node raising
Sluicing
Stripping
VP-ellipsis
Focus particle
Question tag

Exemplary test suite entries De-En

Source	Category	Phenomenon	Target (raw)	Target (edited)	Positive token (indicative)	Negative token (indicative)
Lena machte sich früh vom Acker.	MWE	Idiom	Lena [left the field early].	Lena left early.	left early	field
Lisa hat Lasagne gemacht, sie ist schon im Ofen.	Non-verbal agreement	Coreference	Lisa has made lasagna, [she] is already in the oven.	Lisa has made lasagna, it is already in the oven.	it	she
Ich habe der Frau das Buch gegeben.	Verb tense/ aspect/ mood	Ditransitive - perfect	I [have] the woman of the Book.	I have given the woman the book.	given the book to the woman, gave the book to the woman, given the woman the book, gave the woman the book	

Test suite experiment – systems used

O-PBMT	Old (phrase-based) version of Google Translate (**o**nline, February 2016)
O-NMT	New (neural) version of Google Translate (**o**nline, November 2016)
OS-PBMT	Open-source phrase-based system (Moses) that uses a default configuration to serve as a baseline (only De-En)
DFKI-NMT	Barebone neural system from DFKI, based on an encoder-decoder neural architecture with attention
ED-NMT	Neural system from U Edinburgh, system was built using the Nematus toolkit
RWTH-NMT	NMT-system from RWTH, makes use of subword units and has been finetuned to perform well on the IWSLT 2016 spoken language task (only De-En)
RBMT	Commercial rule-based system Lucy

Test suite experiment – evaluation procedure

QT21

- So far: manual checking
- One phenomenon at a time, e.g.:
  - For ambiguity: Do I find the right sense, no matter what I find in the rest of the sentence?
  - For a prefix verb: Do I find both parts?
  - For an English question: Do I see the Wh-Word and two verbs?
  - For a verb paradigm “X has given Y to Z”: It the sentence complete and correct?
  - …
- Count results

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, Philip Williams. **A Linguistic Evaluation of Rule-based, Phrase-based, and Neural MT Engines.** EAMT 2017, forthcoming

Test suite experiment – results (De-En)

QT21

	#	O-PBMT	O-NMT	RBMT	OS-PBMT	DFKI-NMT	RWTH-NMT	ED-NMT
Ambiguity	17	12%	35%	**42%**	24%	35%	12%	35%
Composition	11	27%	**73%**	55%	27%	45%	45%	**73%**
Coordination & ellipsis	8	25%	**100%**	38%	25%	38%	63%	63%
False friends	5	**40%**	**40%**	20%	20%	20%	**40%**	20%
Function words	19	5%	**68%**	21%	11%	26%	**68%**	42%
LDD & interrogative	66	12%	**79%**	62%	21%	36%	55%	52%
MWE	42	14%	**36%**	7%	21%	10%	12%	19%
NE & terminology	25	48%	48%	40%	**52%**	40%	48%	40%
Negation	6	17%	83%	83%	17%	**100%**	67%	83%
Subordination	36	22%	**58%**	50%	31%	47%	42%	31%
Verb tense/aspect/mood	529	59%	80%	**91%**	52%	53%	74%	63%
Verb valency	32	16%	**50%**	44%	13%	47%	38%	**50%**
Sum	796	363	582	592	341	377	501	446
Average		46%	**73%**	**74%**	43%	47%	63%	56%

Test suite experiment – examples: ambiguity

(1)	Source:	Er hat einen Kater, weil er sehr tierlieb ist.
	Reference:	He has a cat because he is very fond of animals.
	O-PBMT:	He has a hangover, because he is very fond of animals.
	O-NMT:	He has a cat because he is very fond of animals.
	RBMT:	He has a tomcat because it is very animal-dear.
	OS-PBMT:	He has a hangover because it is an encounter.
	DFKI-NMT:	He has a kater because he is very animal.
	RWTH-NMT:	He has a hangover because he‘s very animal.
	ED-NMT:	He has a hangover because he is very animal-loving.

Test suite experiment – examples: phrasal verb

(2)	Source:	Warum hörte Herr Muschler mit dem Streichen auf?
	Reference:	Why did Mr. Muschler stop painting?
	O-PBMT:	Why heard Mr. Muschler on with the strike?
	O-NMT:	Why did Mr. Muschler stop the strike?
	RBMT:	Why did Mr. Muschler stop with the strike?
	OS-PBMT:	Why was Mr Muschler by scrapping on?
	DFKI-NMT:	Why did Mr. Muschler listen to the rich?
	RWTH-NMT:	Why did Mr. Muschler listen to the stroke?
	ED-NMT:	Why did Mr. Muschler stop with the stump?

Test suite experiment – examples: modal particle

QT21

(5)	Source:	Kommst du denn?
	Reference:	Are you coming?
	O-PBMT:	You coming?
	O-NMT:	Are you coming?
	RBMT:	Do you come?
	OS-PBMT:	If you arrive?
	DFKI-NMT:	Do you not?
	RWTH-NMT:	Are you coming?
	ED-NMT:	Are you coming?

Test suite experiment – examples: wh-movement

(6)	Source:	Warum macht der Tourist drei Fotos?
	Reference:	Why does the tourist take three fotos?
	O-PBMT:	Why does the tourist three fotos?
	O-NMT:	Why does the tourist make three fotos?
	RBMT:	Why does the tourist make three fotos?
	OS-PBMT:	Why does the tourist three fotos?
	DFKI-NMT:	Why does the tourist make three fotos?
	RWTH-NMT:	Why is the tourist taking three fotos?
	ED-NMT:	Why does the tourist make three fotos?

Test suite experiment – examples: MWE

(7)	Source:	Die Arbeiter müssten in den sauren Apfel beißen.
	Reference:	The workers would have to bite the bullet.
	O-PBMT:	The workers would have to bite the bullet.
	O-NMT:	The workers would have to bite into the acid apple.
	RBMT:	The workers would have to bite in the acid apple.
	OS-PBMT:	The workers would have to bite the bullet.
	DFKI-NMT:	Workers would have to bite in the acid apple.
	RWTH-NMT:	The workers would have to bite into the clean apple.
	ED-NMT:	The workers would have to bite in the acidic apple.

Test suite experiment – examples: negation

(9)	Source:	Ich glaube, dass es auch nicht die amerikanische Position unterstützt.
	Reference:	I think that it does not support the American position either.
	O-PBMT:	[…] it also does not support the US position.
	O-NMT:	[…] it does not support the American position either.
	RBMT:	[…] it does not support the American position either.
	OS-PBMT:	[…] it is also not the American position.
	DFKI-NMT:	[…] it does not support the American position either.
	RWTH-NMT:	[…] it does not support the American position.
	ED-NMT:	[…] it does not support the American position either.

Test suite experiment – examples: relative clause

(10)	Source:	Wie kann ich die Farbe, mit der ich arbeite, ändern?
	Reference:	How can I change the color I am working with?
	O-PBMT:	How can I change the color with which I work to change?
	O-NMT:	How can I change the color with which I work?
	RBMT:	How can I change the color with which I work?
	OS-PBMT:	How can I change the colour, with whom i work, change?
	DFKI-NMT:	How can I change the color I work with?
	RWTH-NMT:	How can I change the color I work with?
	ED-NMT:	How can I change the color I work with?

Test suite experiment – examples: modal negated pluperfect subjunctive II

(11)

Source:	Ich hätte nicht lesen gedurft.
Reference:	I would not have been allowed to read.
O-PBMT:	I would not have been allowed to read.
O-NMT:	I should not have read.
RBMT:	I would not have been allowed to read.
OS-PBMT:	I would not have read gedurft.
DFKI-NMT:	I would not have been able to read.
RWTH-NMT:	I wouldn‘t have read.
ED-NMT:	I wouldn‘t have read.

Test suite experiment – examples: case government

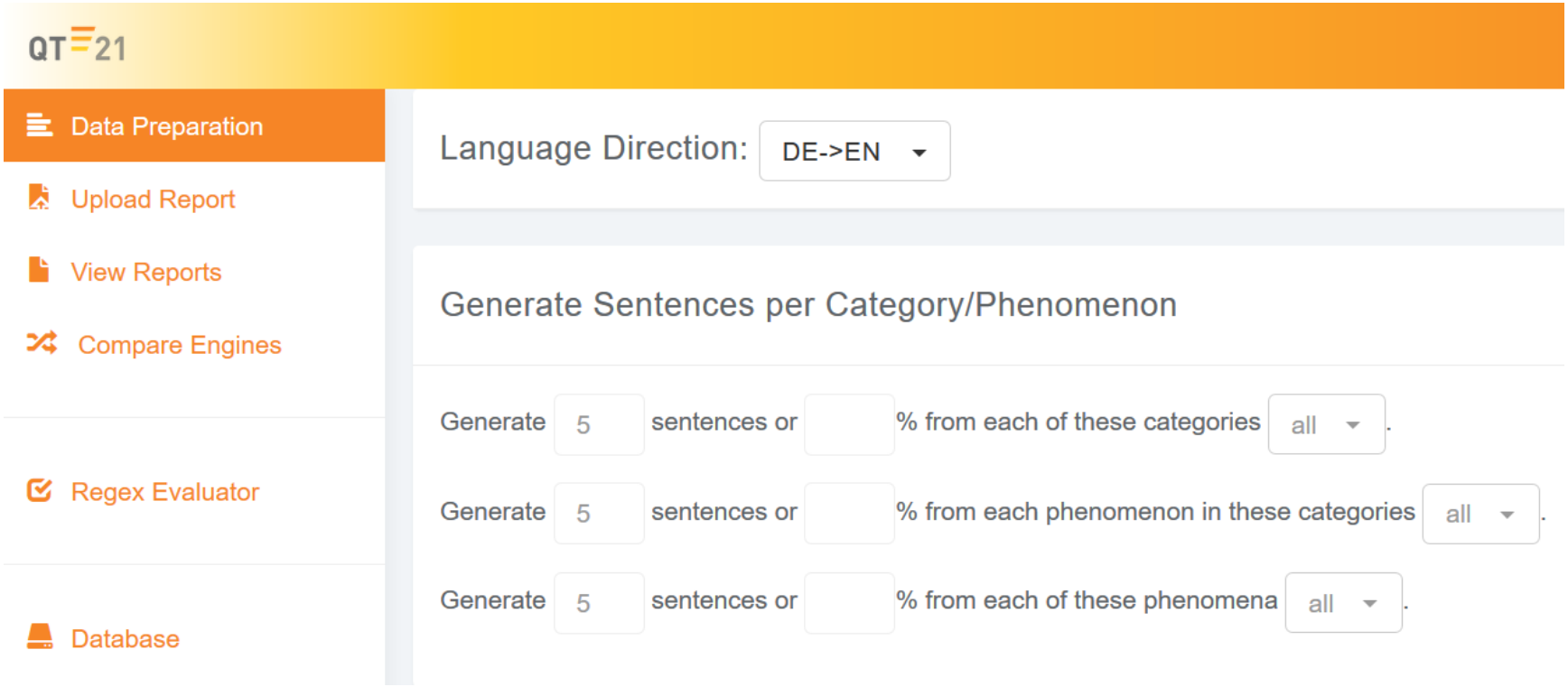
(12) Source: Der Manager besteht auf den Test.
Reference: The manager insists on the test.

O-PBMT: The manager is on the test.
O-NMT: The manager insists on the test.
RBMT: The manager insists on the test.
OS-PBMT: The manager is on the test.
DFKI-NMT: The manager is on the test.
RWTH-NMT: The manager is on the test.
ED-NMT: The manager is on the test.

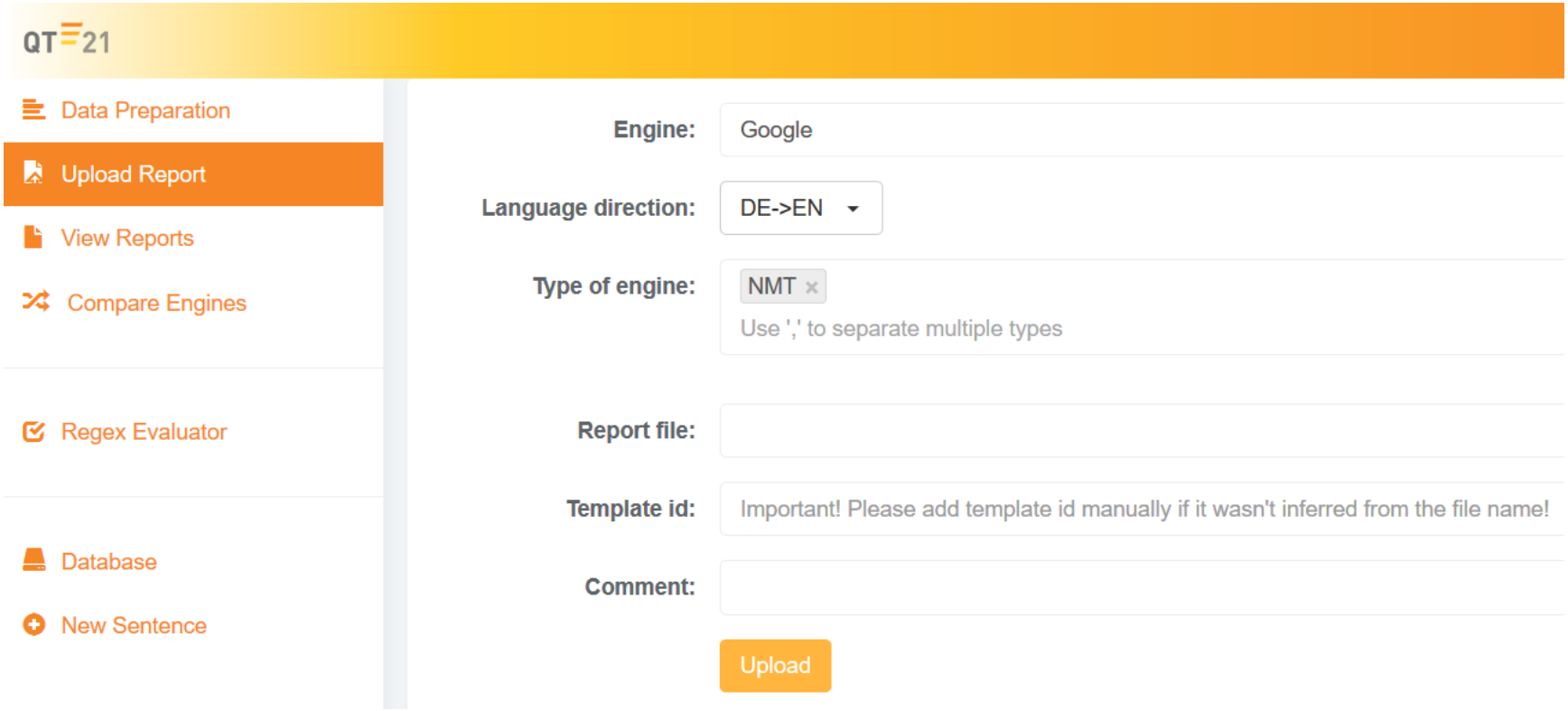
QT21

TEST SUITE AUTOMATION

Data preparation



Report upload

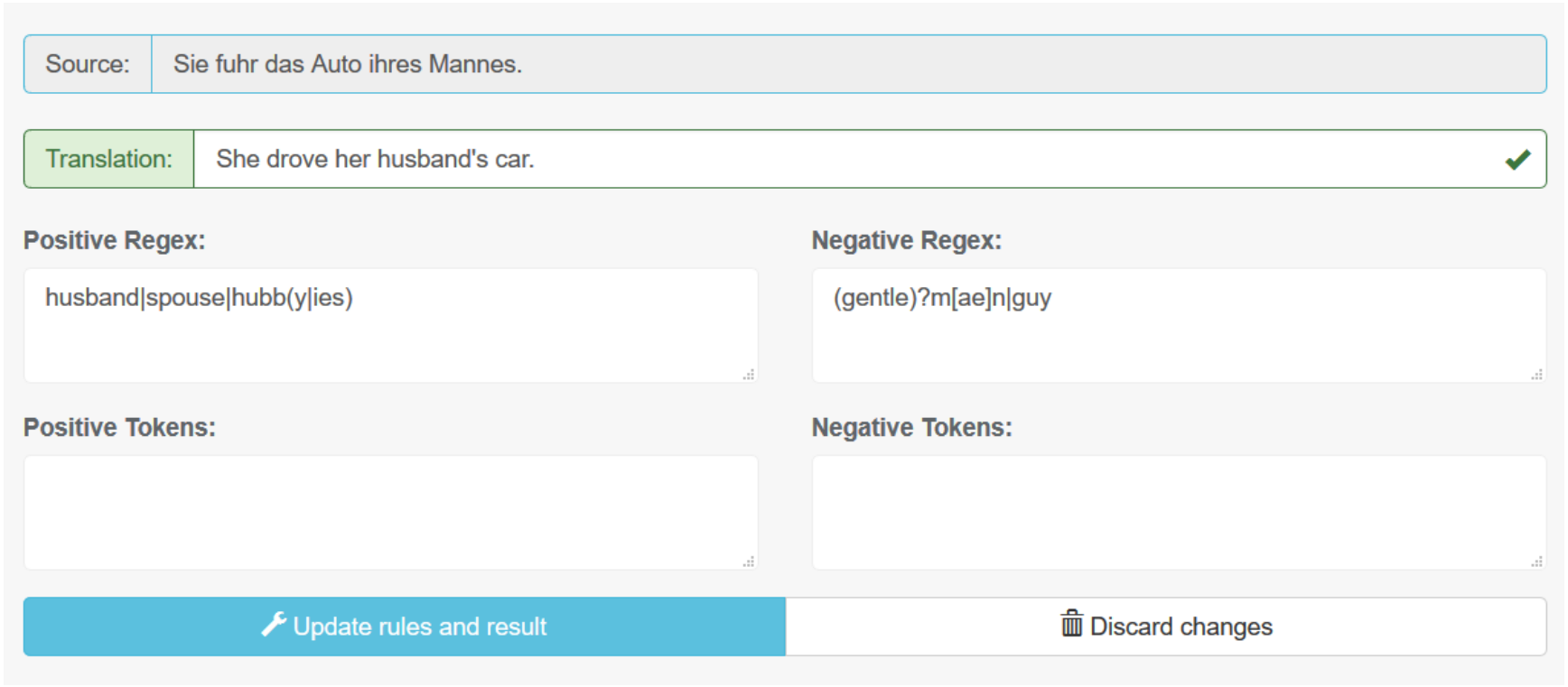


Evaluation

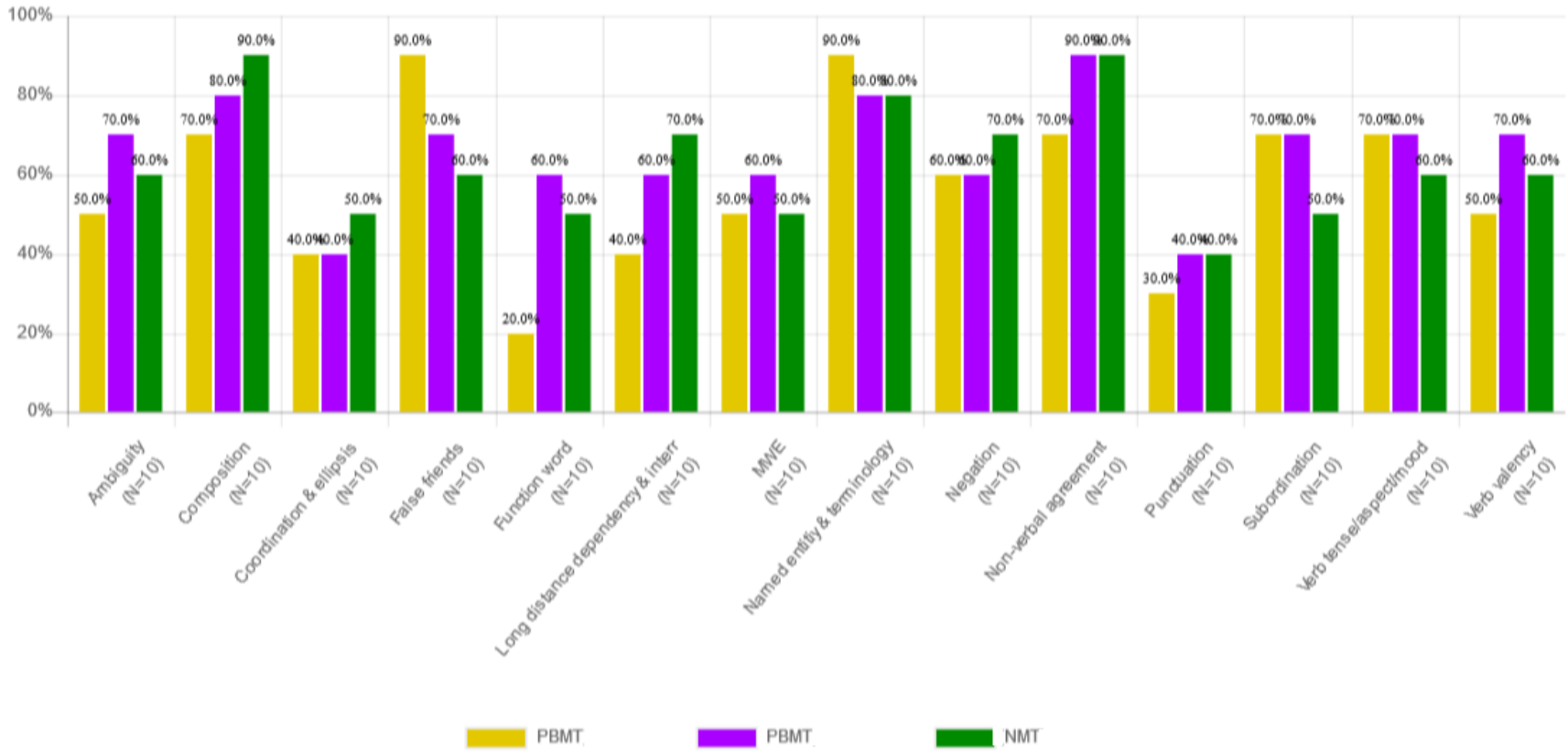
Apply Tokens 0 | Skip Tokens

	ID data point	Source	Category	Phenomenon	Translation
⚠	00005001	Er hob die Tür aus den Angeln.	Ambiguity	Lexical ambiguity	He lifted the door.
✔	00005002	Die Tür hing schief in den Angeln.	Ambiguity	Lexical ambiguity	The door hung obliquely in the hinges.
✖	00005003	Die Angeln der Tür quietschten beim Öffnen.	Ambiguity	Lexical ambiguity	The angling of the door creaked open.

Regular Expressions



Comparison



LREC 2018 paper forthcoming

CUSTOM TEST SUITES

Technical test suite example

	#	PB-SMT	RBMT	RBMT improved	neural	sel. mech.
imperatives	247	68%	79%	**79%**	74%	*73%
compounds	219	55%	87%	**85%**	51%	70%
“>” separators	148	**99%**	39%	83%	93%	80%
quotation marks	431	**97%**	94%	75%	95%	80%
verbs	505	85%	93%	**93%**	90%	*90%
phrasal verbs	90	22%	68%	**77%**	38%	53%
terminology	465	**64%**	50%	53%	55%	54%
sum	2105					
average		76%	77%	77%	75%	74%

Eleftherios Avramidis, Vivien Macketanz, Aljoscha Burchardt, Jindrich Helcl and Hans Uszkoreit “Deeper Machine Translation and Evaluation for German”. DMTW 2016

Recent study on customer data

QT21

- Adopted Moses vs. unadopted NMT

	#	NMT	Moses
formal address	138	90%	86%
genitive	114	**92%**	68%
modal construction	290	**94%**	75%
negation	101	93%	86%
passive voice	109	**83%**	40%
predicate adjective	122	81%	75%
prepositional phrase	104	81%	75%
terminology	330	35%	**68%**
tagging	145	83%	**100%**
sum	1453		
average		**89%**	73%

Table 2: Manual evaluation translation accuracy focusing on particular phenomena.

Anne Beyer, Vivien Macketanz, Aljoscha Burchardt and Philip Williams. **Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data?** EAMT 2017, forthcoming

Conclusions

- Current evaluation workflow based on reference translation (and scores like BLEU) provides little insights about MT quality and the nature of errors
- Alternatives are being actively researched:
  - Learning from post-edits
  - Target analytics: Error annotation with MQM
  - Task-based evaluation
  - Source-driven testing: Test suites
  - Quality estimation, better automatic metrics, etc.
- Still: communication between communities (MT development and language experts) can be intensified

Thank you

Quality Translation 21 (QT21) has received funding from the EU's Horizon 2020 research and innovation programme under grant no. 645452. META-QT has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357). Formerly co-funded by FP7 and ICT PSP through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899).