



A New Deal for Translation Quality

Aljoscha Burchardt

Fellowship & Colloquial Talk

DFKI Berlin, 15.10.2020

First things first: Why do I like machine translation?

Google Übersetzer



The screenshot shows the Google Translate web interface. At the top, there are two tabs: 'Text' (selected) and 'Dokumente'. Below the tabs, there are language selection options: 'SPRACHE ERKENNEN', 'TÜRKISCH', 'DEUTSCH' (selected), and 'ENGLISCH'. On the right side, there are more language options: 'DEUTSCH', 'ENGLISCH' (selected), and 'FRANZÖSISCH'. The main input area contains the German text 'der, die, das, wieso, weshalb, warum' and the output area contains the English text 'the, the, the, why, why, why'. There are also icons for voice input/output and a character count '36/5000'.

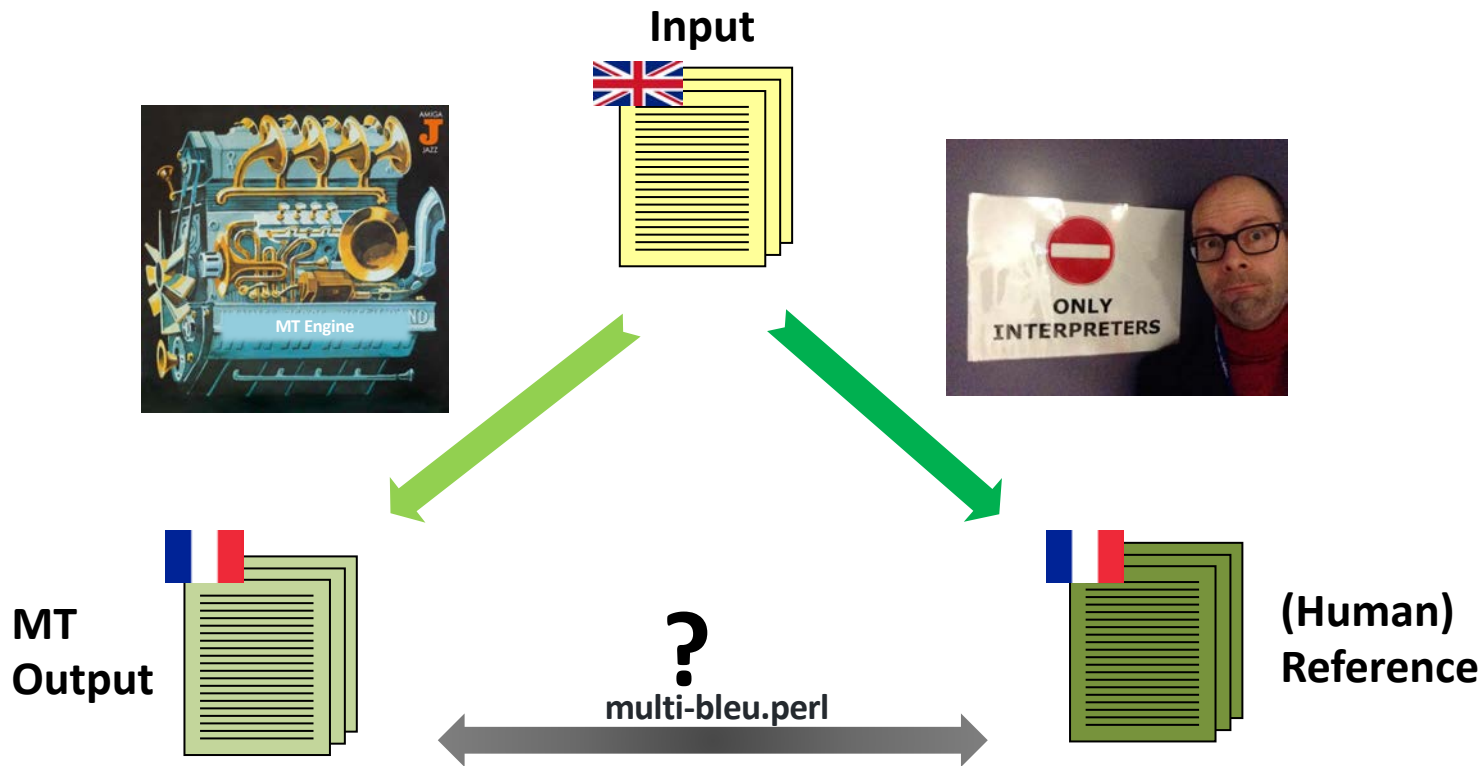


Image: Sesame Workshop

How it all started ... back in 2010



Assessing quality in MT development





- In many NLP tasks, performance can be measured as deviation from some ideal (POS tagging, fact extraction, etc.)
- In MT, this is difficult
 - Theoretical issue: there is **no eternal notion of “good translation”**, MT quality is task-specific: length, target audience, style guide, etc.
 - 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.*

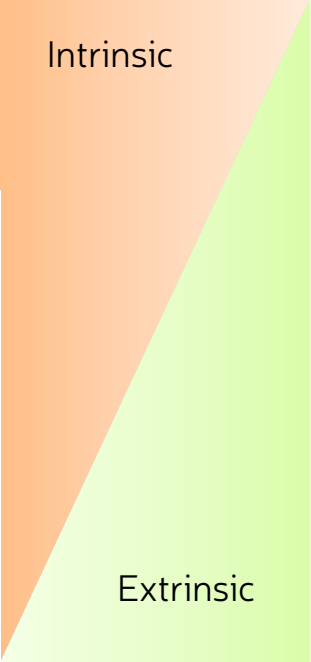
Reference-based assessment of MT output



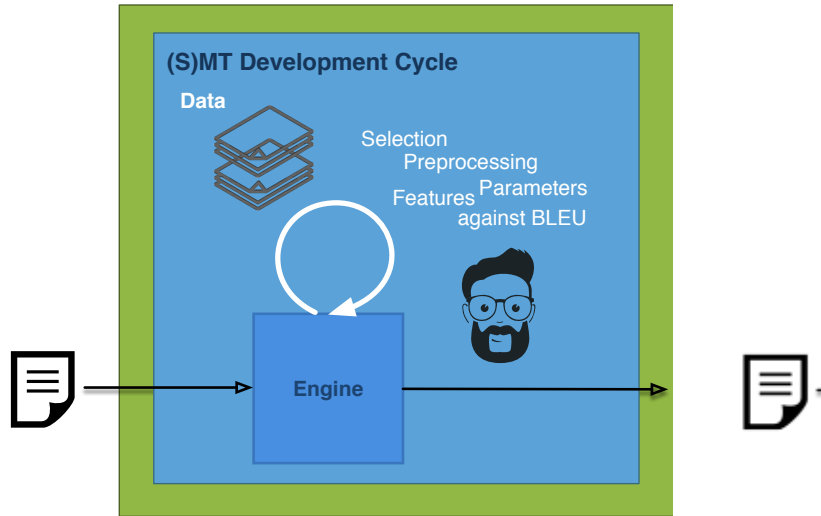
	Reference-based approach (BLEU etc.)
Assessment of single segments	X
Shows number of errors	X
Shows type of errors	X
Meaningful comparison between different systems, system types, languages	X
Applicable w/o human reference translation	X
Automatic	✓

Who needs MT-Evaluation?



	Means	Task-specific?
<ul style="list-style-type: none">• MT Researchers:<ul style="list-style-type: none">• Rapid feedback for engineering.• Which setting is better?• Are differences significant?	Shallow surface comparison with one (!) reference translation	 <p>Intrinsic</p> <p>Extrinsic</p>

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

How humans can provide feedback

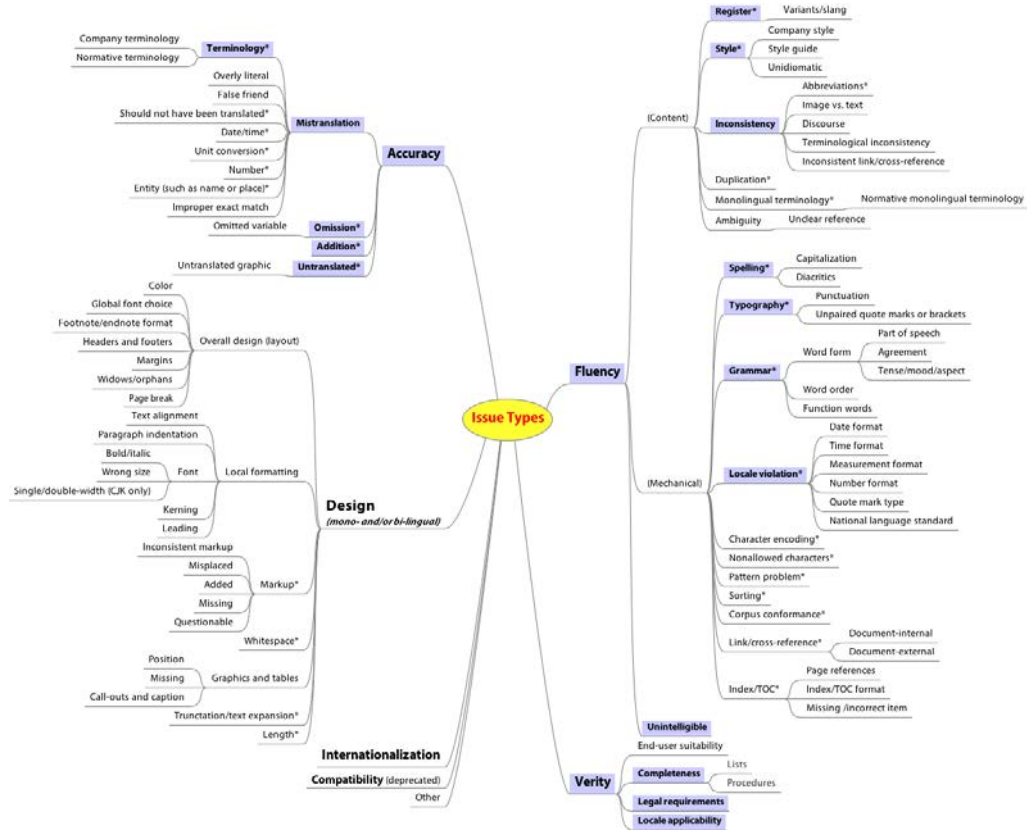


- Analytic error annotation (MQM)
- Task-based evaluation
- Designing test suites

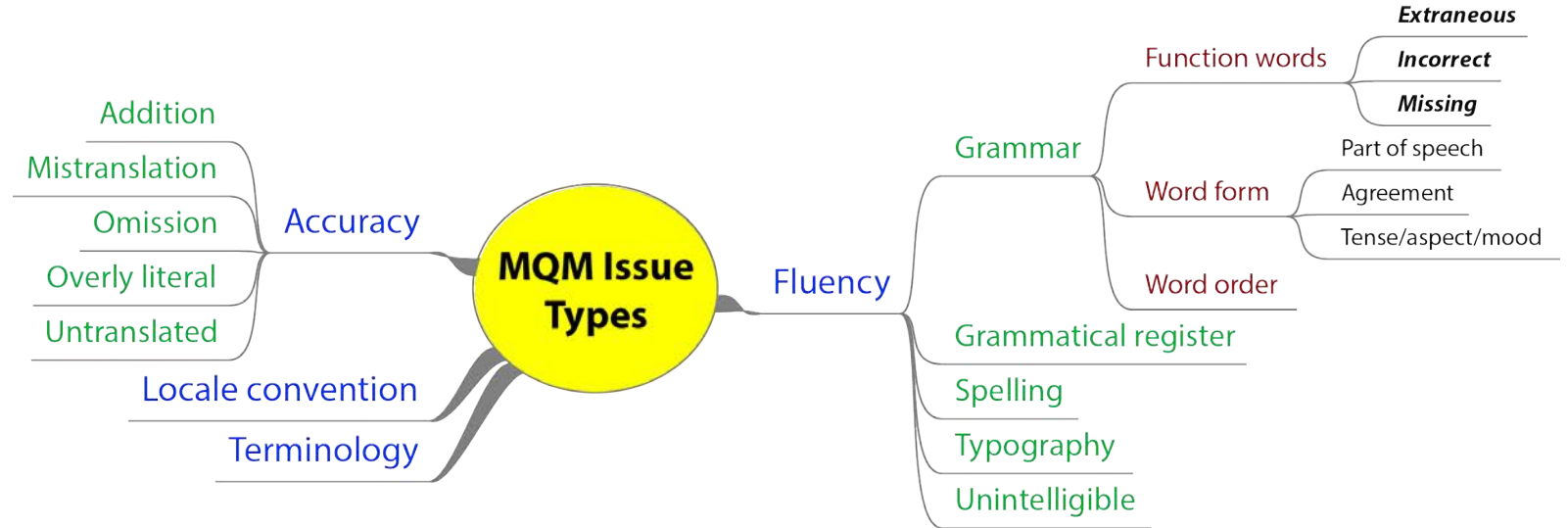


MULTIDIMENSIONAL QUALITY METRICS (MQM)

Full MQM Structure



MT evaluation metric in QT21



- Includes three custom issues
- Heavily tuned to address Grammar issues

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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 1. Mistranslation [Sprung zu] 2. Typography [.] 3. Omission []

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

Arlé 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*

Aljoscha Burchardt, Arlé Richard Lommel, Lindsay Bywood, Kimberley Harris, Maja Popovic
Machine translation quality in an audiovisual context in: Yves Gambier, Sara Ramos Pinto (eds.): *Target volume 28 number 2, Pages 206-221, John Benjamins, 2016*

Error profiles by system and language



Error type	DE-EN	EN-DE		EN-LV		EN-CS
	PBMT	PBMT	NMT	PBMT	NMT	PBMT
Accuracy	3	0	0	39	50	0
Addition	539	332	167	277	268	385
Mistranslation	437	967	852	274	677	786
Omission	576	690	355	295	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	3336	6775	3700	3803	3635	8321

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

Update



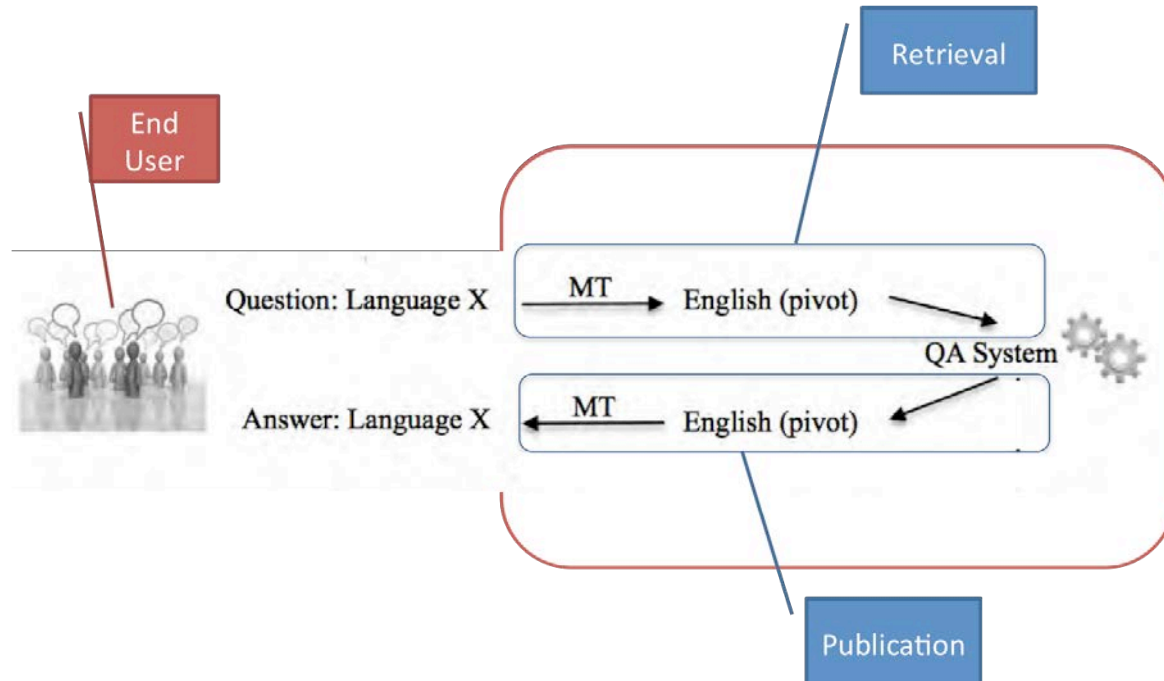
- MQM is currently in the standardization process in ASTM Committee F43 (American Society for Testing and Materials; now known as ASTM International)

TASK-BASED EVALUATION

Extrinsic Evaluation Scenario



Question: How probable is it that the human operator is being called?



Pilot 0: Emulate Real Usage



0. A questão:

No meu iPad está a aparecer uma seta virada para a direita, o que é?

0. A resposta A que acabou de ler:

Significa uma candidatura é utilizar serviços de localização.

1. Agora leia a resposta B:

Quer dizer que um aplicativo está usando os Serviços de Localização.

2. Considerando que a resposta B dá a informação correcta, qual do seguinte é verdade acerca da resposta A:

A dá o conselho correcto.

A tem pontos menores errados.

A tem pontos importantes errados.

- 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.

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%



TEST SUITES

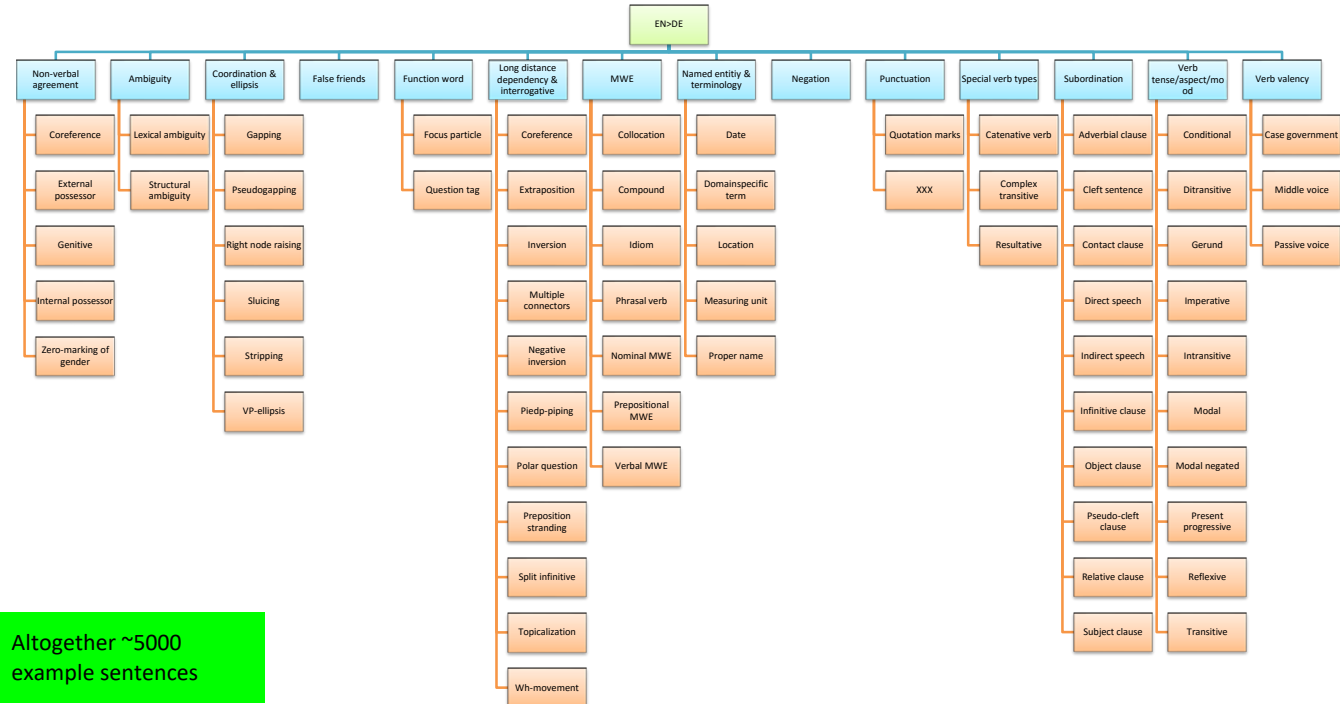


- Test suites are a familiar tool in NLP in areas such as grammar development.
- Idea: Use test suites in MT development.
- Systematically evaluate and compare system(variant)s
 - Gets all 20 imperatives right
 - Gets half of the imperatives right
 - Gets no imperatives right
 - ...
- Guide system improvement / 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...

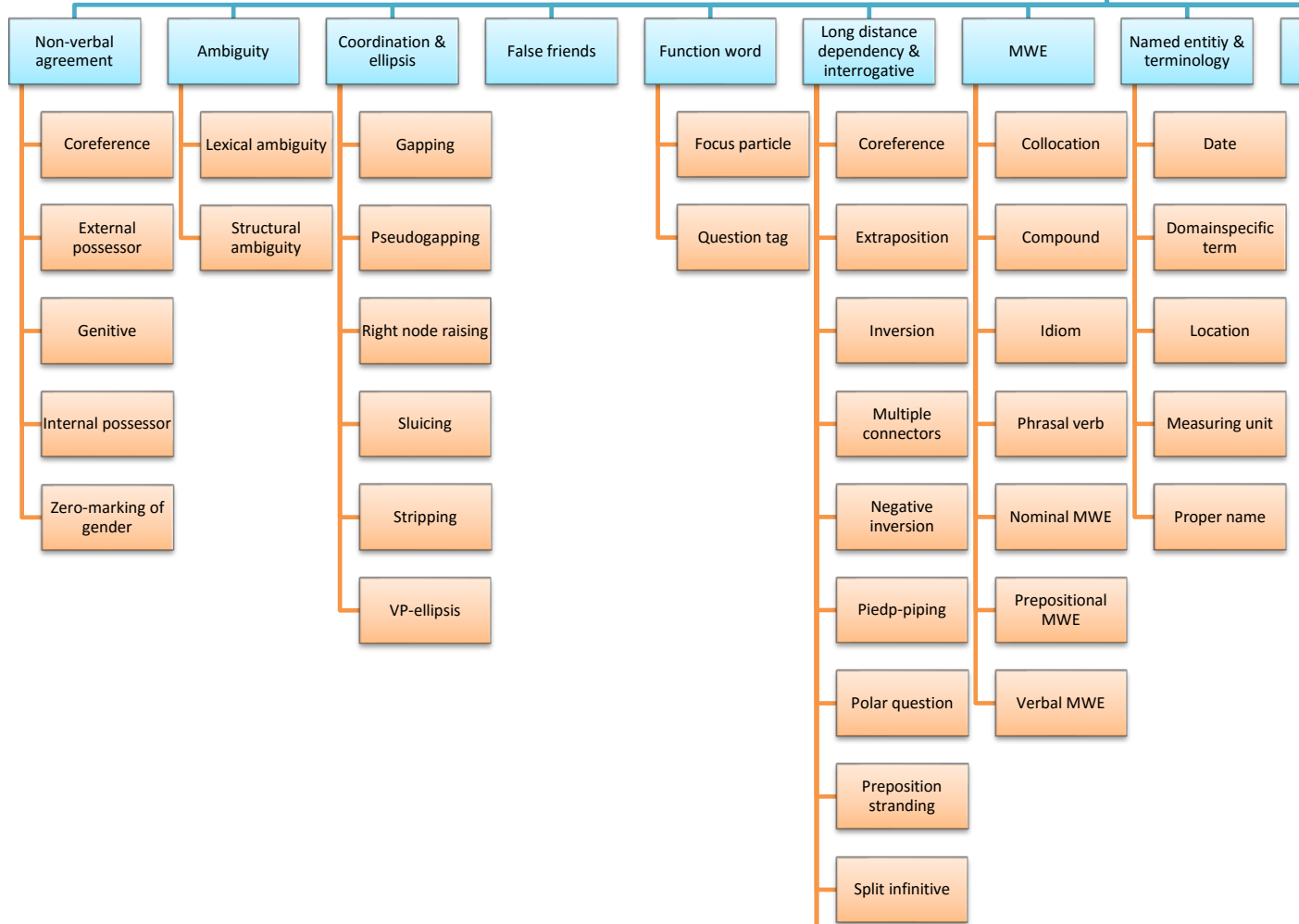
14 Barrier Categories

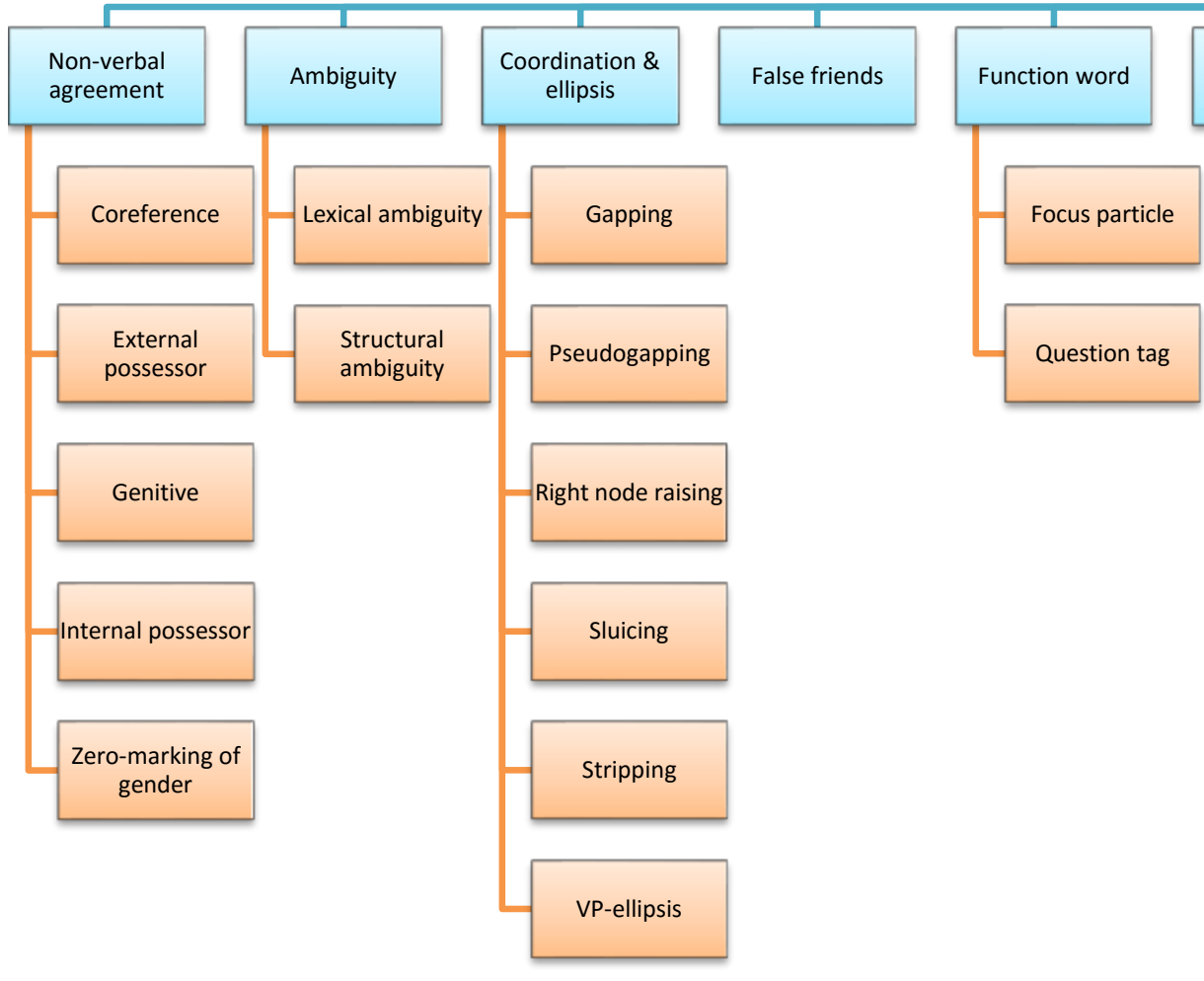


~ 65 Barriers



Altogether ~5000 example sentences





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 (online, February 2016)
- O-NMT New (neural) version of Google Translate (online, 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 – 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.

Google-old: He has a hangover, because he is very fond of animals.

Google-new: 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?
- Google-old: Why heard Mr. Muschler on with the strike?
Google-new: Why did Mr. Muschler stop the strike?
Update 2020: Why did Mr. Muschler stop painting?
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: MWE





- (7) Source: Die Arbeiter müssten in den sauren Apfel beißen.
Reference: The workers would have to bite the bullet.

Google-old: The workers would have to bite the bullet.
Google-new: The workers would have to bite into the acid apple.
Update 2020: The workers would have to bite the bullet.
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 AUTOMATION




Regular Expressions

Source:	Sie fuhr das Auto ihres Mannes.	
Translation:	She drove her husband's car.	✓
Positive Regex:	<code>husband spouse hubb(y ies)</code>	Negative Regex:
		<code>(gentle)?m[ae]n guy</code>
Positive Tokens:		Negative Tokens:
 Update rules and result		 Discard changes

Semi-automatic 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.

Update & Thanks

[Eleftherios Avramidis](#), [Vivien Macketanz](#), [Ursula Strohriegel](#), [Aljoscha Burchardt](#), [Sebastian Möller](#)

[Fine-grained linguistic evaluation for state-of-the-art Machine Translation](#)

In: Proceedings of the Fifth Conference on Machine Translation (WMT-2020)

[Eleftherios Avramidis](#), [Vivien Macketanz](#), [Ursula Strohriegel](#), [Hans Uszkoreit](#)

[Linguistic evaluation of German-English Machine Translation using a Test Suite](#)

In: Proceedings of the Fourth Conference on Machine Translation (WMT-2019)

[Vivien Macketanz](#), [Eleftherios Avramidis](#), [Aljoscha Burchardt](#), [Hans Uszkoreit](#)

[Fine-grained evaluation of German-English Machine Translation based on a Test Suite](#)

In: Proceedings of the Third Conference on Machine Translation (WMT-2018)

category	#	JHU	MLLP	onlA	onlB	onlG	onlY	RWTH	UCAM	UEDIN	avg
Ambiguity	74	-2.7	21.6	4.1	0.0	4.1	10.8	-1.3	2.7	12.1	6.9
Composition	42	4.8	0.0	14.3	0.0	9.5	2.4	-2.4	-4.7	7.1	5.2
Coordination and ellipsis	23	8.7	-4.4	0.0	0.0	13.1	0.0	0.0	-13.1	0.0	7.3
False friends	34	-3.0	5.8	0.0	3.0	-5.9	23.6	5.9	-5.8	14.7	6.8
Function word	41	-2.5	7.3	4.9	0.0	41.4	0.0	-7.4	-2.4	9.7	12.5
LDD & interrogatives	38	10.6	10.6	-2.7	0.0	5.3	0.0	0.0	5.3	7.9	5.6
MWE	53	5.6	7.5	5.7	0.0	1.9	1.9	3.8	-1.8	3.8	4.7
Named entity and terminology	34	5.9	3.0	5.9	0.0	-3.0	-5.9	8.9	0.0	5.9	0.3
Negation	19	0.0	0.0	0.0	0.0	42.1	0.0	0.0	0.0	-10.5	6.6
Non-verbal agreement	48	12.5	10.4	12.5	0.0	22.9	2.1	-2.1	0.0	12.5	9.7
Punctuation	51	5.9	2.0	-21.6	0.0	-7.9	1.9	27.5	0.0	23.5	8.0
Subordination	31	3.3	6.5	-6.5	3.2	19.4	3.2	6.5	0.0	0.0	5.0
Verb tense/aspect/mood	3995	-4.0	-5.9	12.9	0.2	19.8	1.6	5.6	-7.6	5.1	6.0
Verb valency	30	10.0	0.0	0.0	0.0	13.4	6.6	0.0	0.0	3.4	5.8
average (items)	4513	-3.1	-4.3	11.6	0.2	18.7	2.0	5.3	-6.8	5.4	6.1
average (categories)		3.9	4.6	2.1	0.5	12.6	3.4	3.2	-2.0	6.8	6.5

Table 2: Percentage (%) of accuracy improvement or deterioration between WMT18 and WMT19 for all the systems submitted (averaged in last column) and the systems submitted with the same name

That's it



Dr. Aljoscha Burchardt
DFKI, Language Technology Lab
Alt-Moabit 91c, 10559 Berlin
Aljoscha.Burchardt@dfki.de
Tel. +49-30-23895-1838

