Statistical Machine Translation LING-462/COSC-482 Week 12: Computer Aided Translation

Achim Ruopp
achim.ruopp@Georgetown.edu

Agenda

- Language in ten minutes: Hawaiian – Austin Blodgett
- Quality Goals
- Human Translation Process and Quality Assurance
- Machine Translation Post-Editing

- Break -

- Interactive Translation Prediction
- User Studies

Researchers and Users have dramatically different **Goals**

MT in the lab: Build Autonomous Translation Machines

MT in practice:
Integrate MT with workflow and human expertise

Note: This is an accurate representation of a real translation workflow ;)

Mike Dillinger, "MT escaped from the Lab!", Keynote AMTA 2016

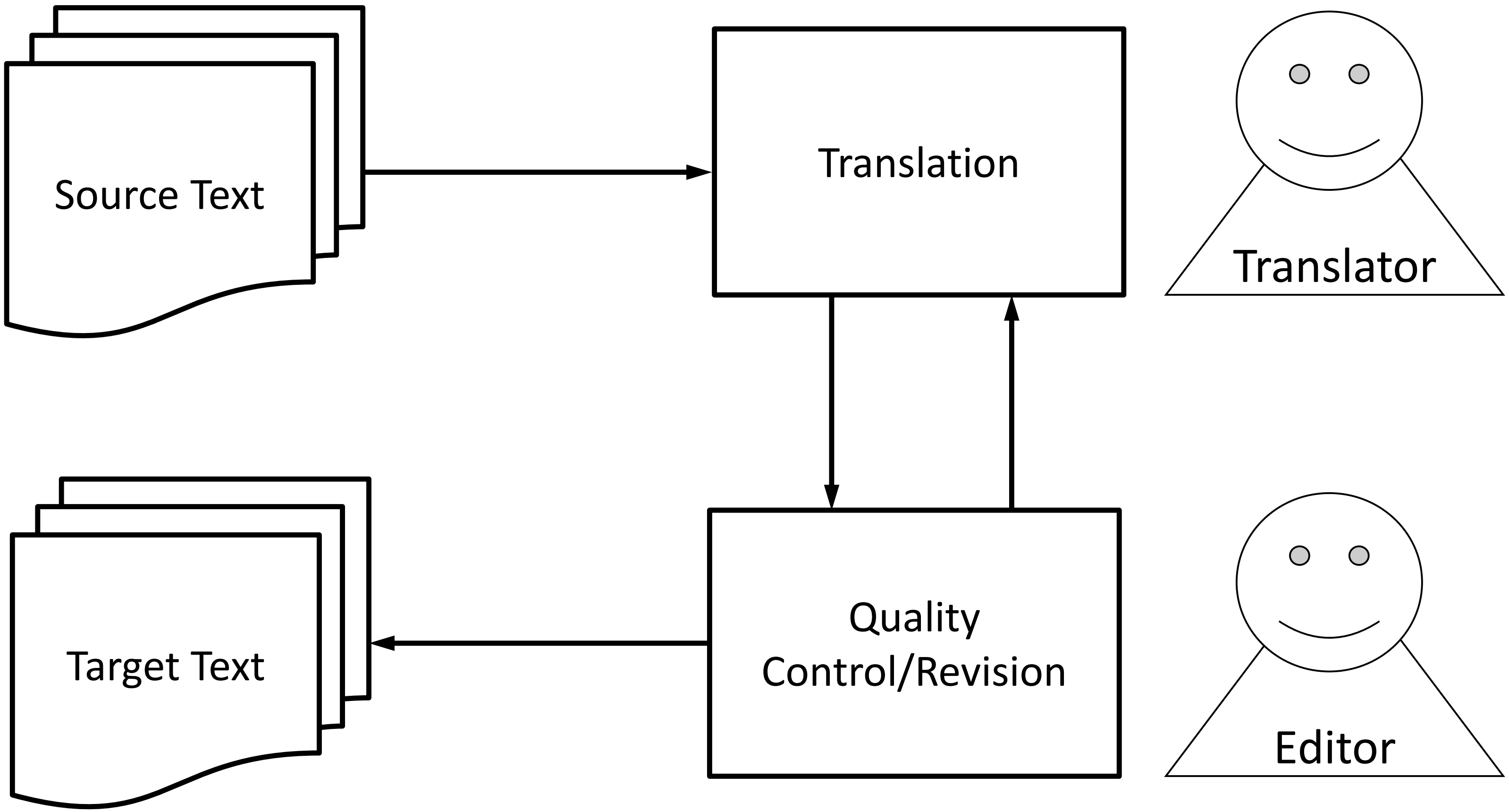
Quality Goals for Translation

- A lot of discussion around this
- But people usually can only agree on two levels
    1. High-quality human translation and revision/publishable quality
    2. Good enough/fit for purpose
- Level 1. is usually what people are willing to pay good money for (e.g. $0.10/word)
- Level 2. is for high-volume, less important, perishable content

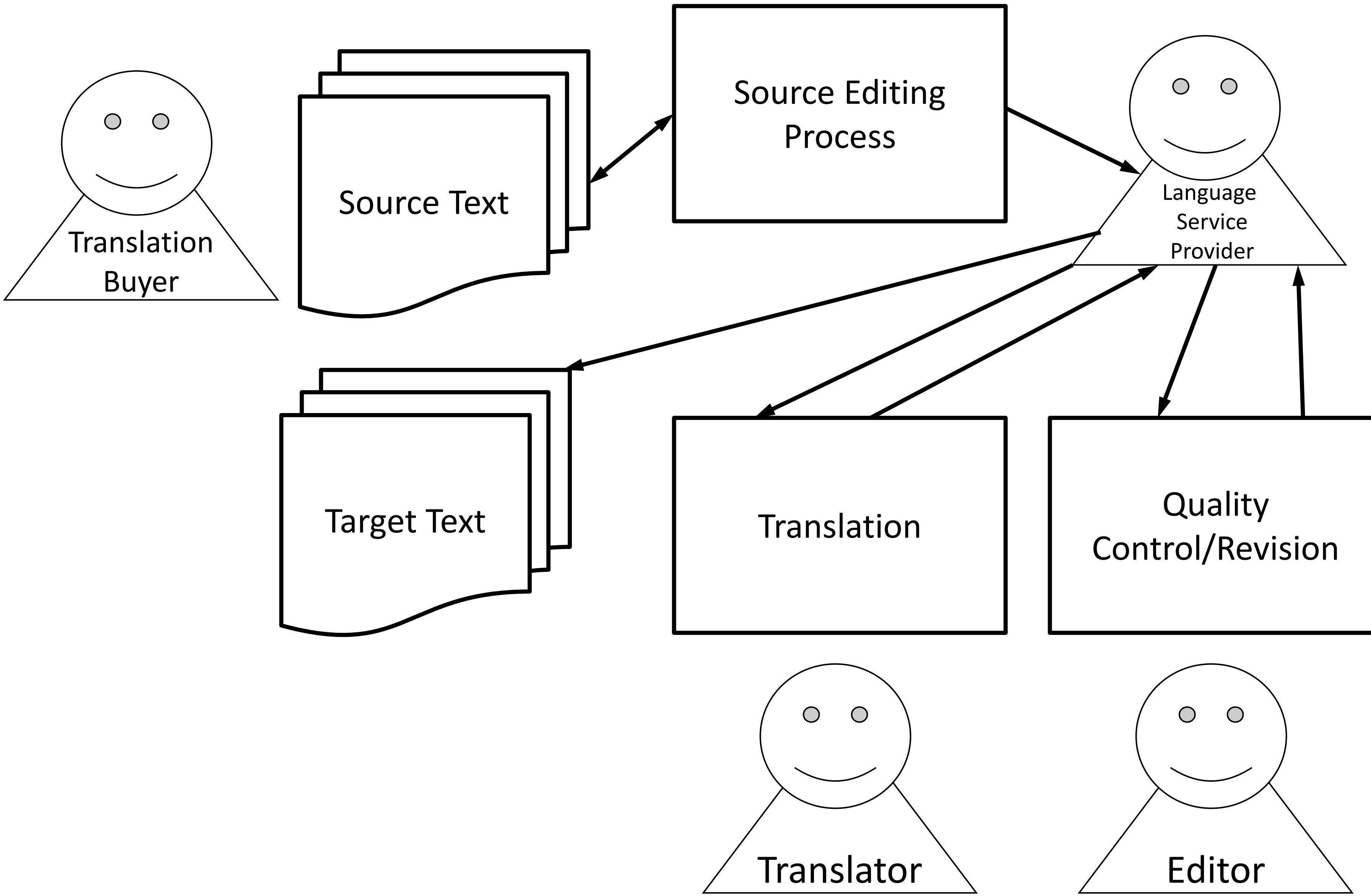
Level 2: Good enough/Fit for Purpose Quality

- Usually for gisting
- Does not impact reputation of publisher
  - At least when properly disclosed
- Increasingly MT-only
  - No human translator/editor in the loop
- Andy Way. 2018. Quality Expectations of Machine Translation . Translation Quality Assessment: From Principles to Practice, Springer.

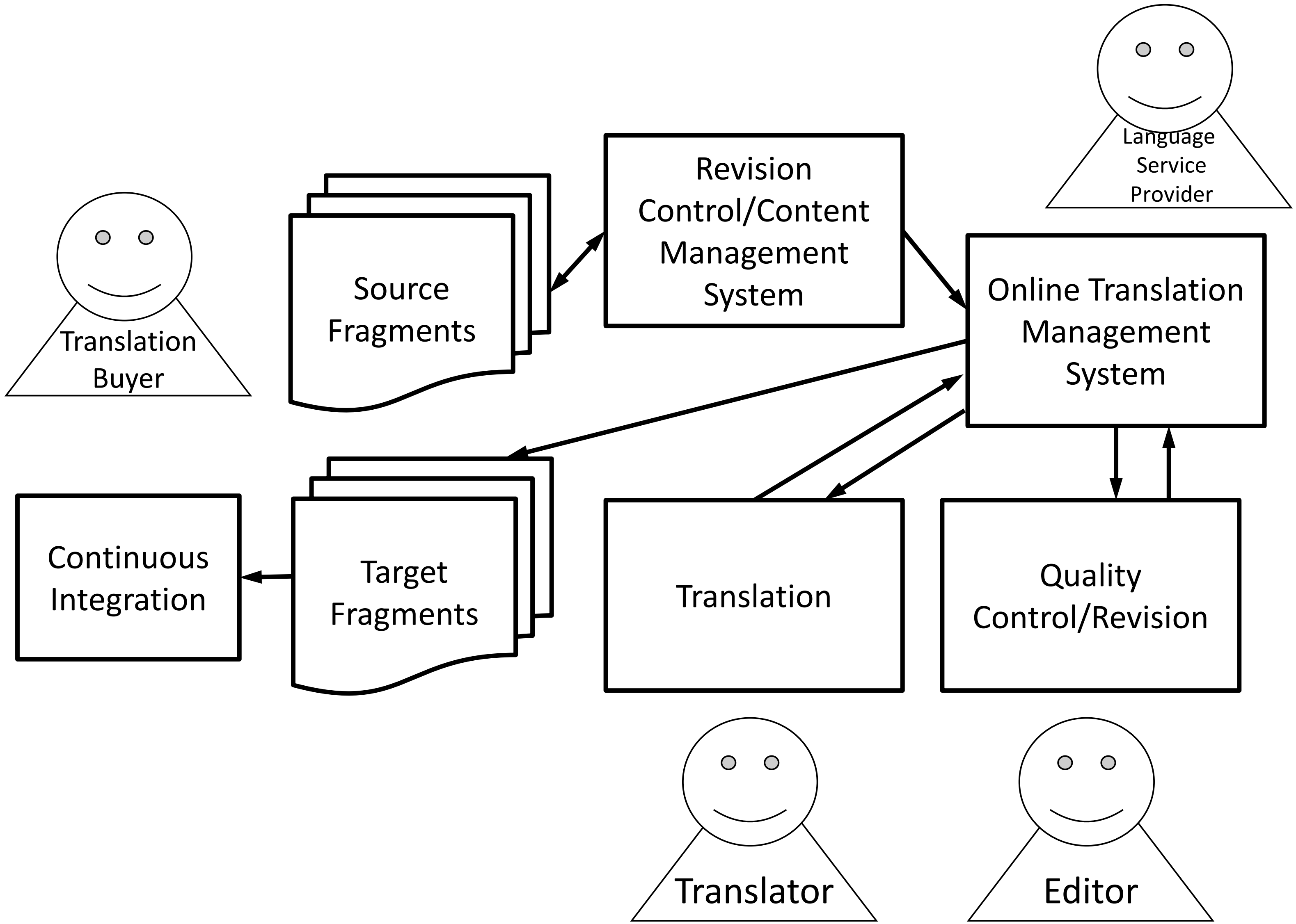
“Waterfall” Process for High-Quality Human Translation and Revision



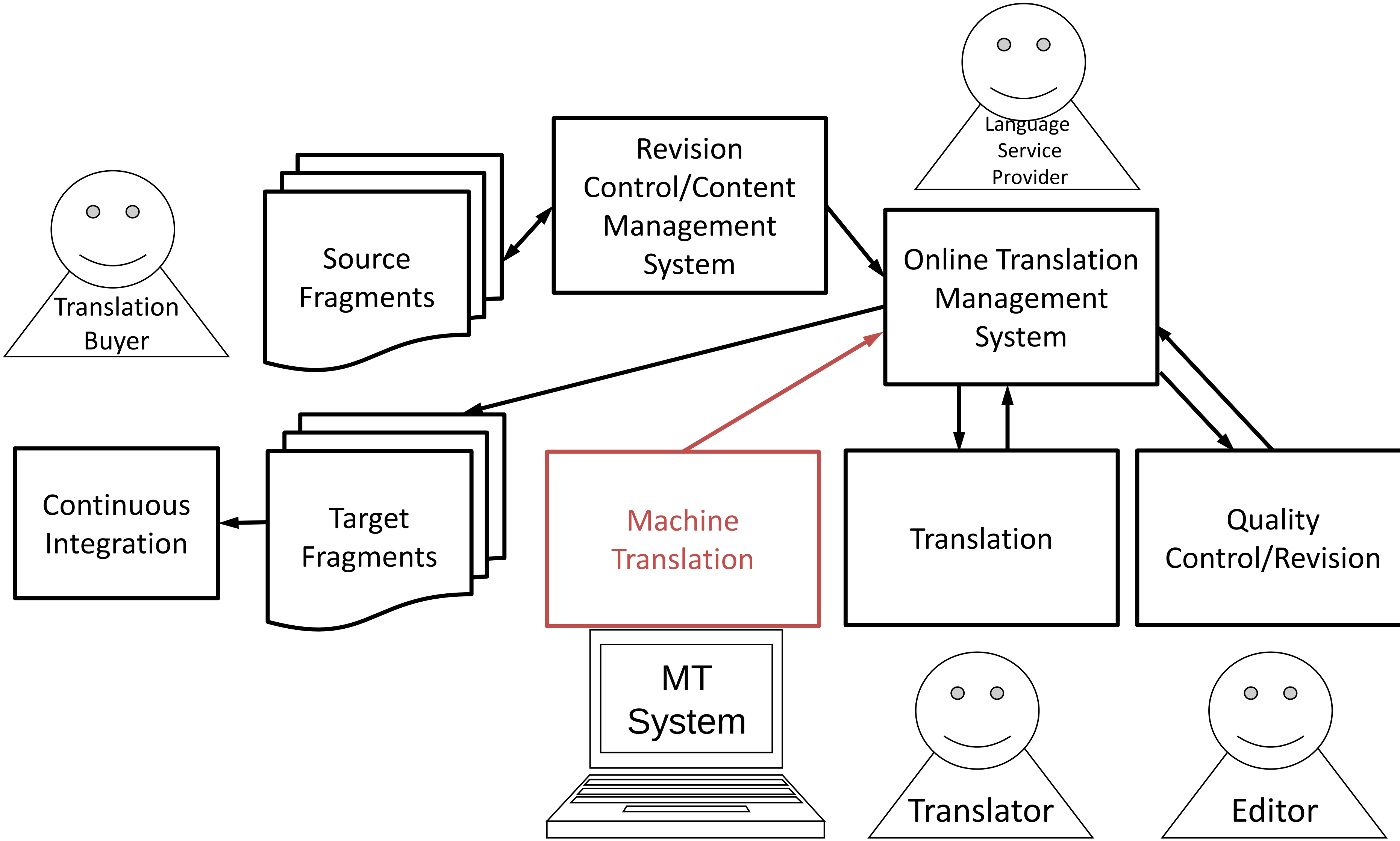
“Waterfall” Process in Practice



Agile Translation Process



Agile Translation Process with Machine Translation Post-Editing



What Toolkits don’t include for MT integration

- Support for all needed languages
- Pre-built MT systems
- Graphical User Interface/Web UI
- Process/API integration into CAT tools
- Handling of formatting
- Handling of terminology
- Only limited: Adaptive and interactive MT
- Test suite
- Supported releases

TAUS/DCU Post-editing guidelines for high-quality translation

- Aim for grammatically, syntactically and semantically correct translation.
- Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms".
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling, punctuation and hyphenation apply.
- Ensure that formatting is correct.

TAUS/DCU Post-editing guidelines for “good enough” translation

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling apply.
- No need to implement corrections that are of a stylistic nature only.
- No need to restructure sentences solely to improve the natural flow of the text.

OMEGAT POST-EDITING DEMO

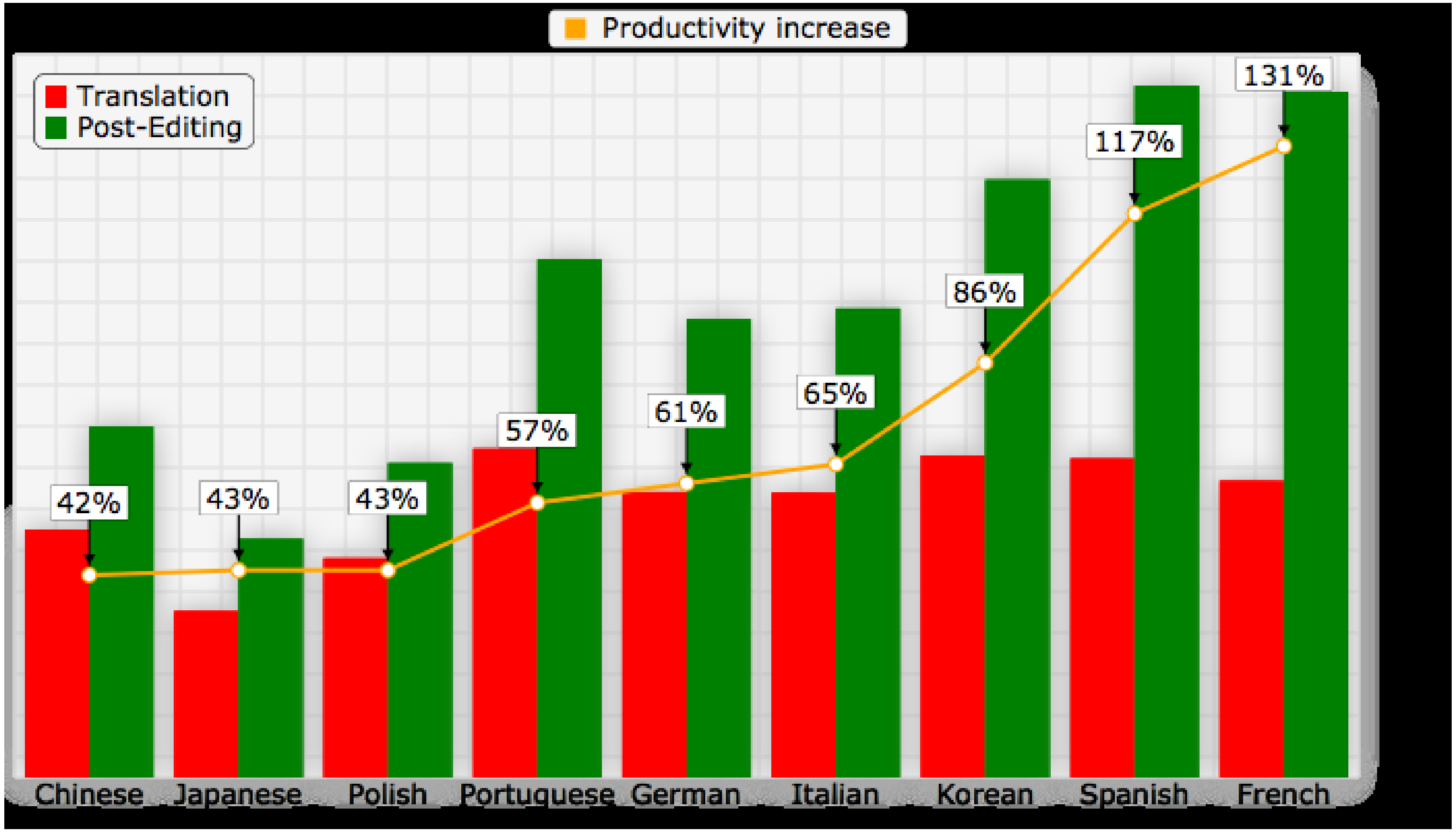
Adaptive MT and Online Learning

- Adaptive MT
  - Adapts to source text domain
  - Adapts to provided translation memory
- Online Learning (aka interactive MT)
  - Live adaptation to post-edits
  - Won't repeat same errors – hopefully!
  - Can be applied to generic or adapted model
- Can improve MT post-editing/reduce frustrations

Machine Translation Post-Editing

- Good news
  - 20%-40% increased translation speed in many high-volume translation scenarios e.g. technical manuals
- Bad news
  - Translators don’t like it
  - Pricing model is unclear (per word? per time?)
  - Is not Hybrid-intelligence translation (Dillinger)
    - Humans + machines working together
    - No hybridization - two separate intelligences
    - No strategic use of best of both
    - No real-time interaction

Post-Editing Machine Translation



(source: Autodesk)

MT Quality and Productivity

System	BLEU	Training Sentences	Training Words (English)
MT1	30.37	14,700k	385m
MT2	30.08	7,350k	192m
MT3	29.60	3,675k	96m
MT4	29.16	1,837k	48m
MT5	28.61	918k	24m
MT6	27.89	459k	12m
MT7	26.93	230k	6.0m
MT8	26.14	115k	3.0m
MT9	24.85	57k	1.5m

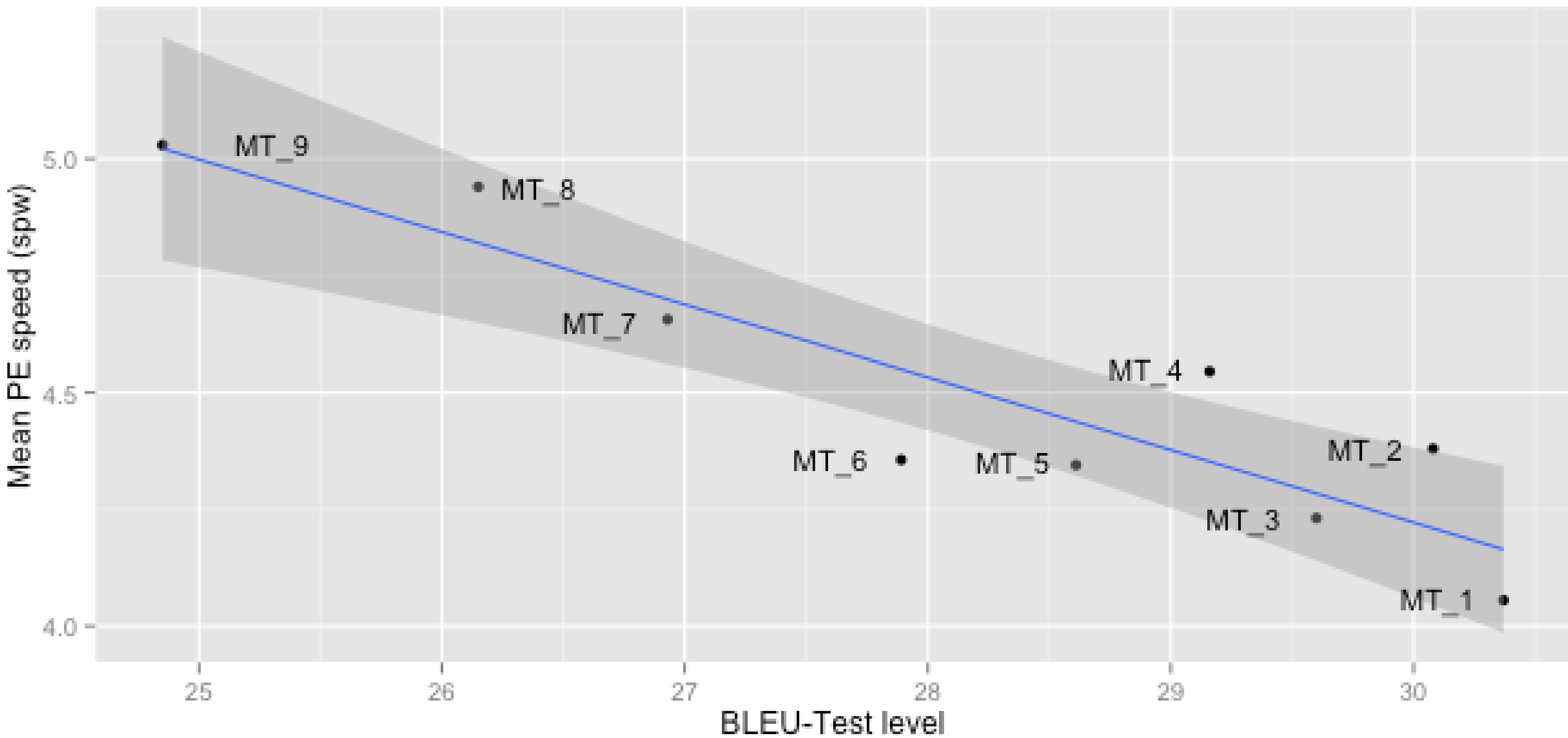
- Same type of system (Spanish–English, phrase-based, Moses)
- Trained on varying amounts of data [Sanchez-Torron and Koehn, AMTA 2016]

MT Quality and Productivity

System	BLEU	Training Sentences	Training Words (English)	Post-Editing Speed
MT1	30.37	14,700k	385m	4.06 sec/word
MT2	30.08	7,350k	192m	4.38 sec/word
MT3	29.60	3,675k	96m	4.23 sec/word
MT4	29.16	1,837k	48m	4.54 sec/word
MT5	28.61	918k	24m	4.35 sec/word
MT6	27.89	459k	12m	4.36 sec/word
MT7	26.93	230k	6.0m	4.66 sec/word
MT8	26.14	115k	3.0m	4.94 sec/word
MT9	24.85	57k	1.5m	5.03 sec/word

- User study with professional translators
- Correlation between BLEU and post-editing speed?

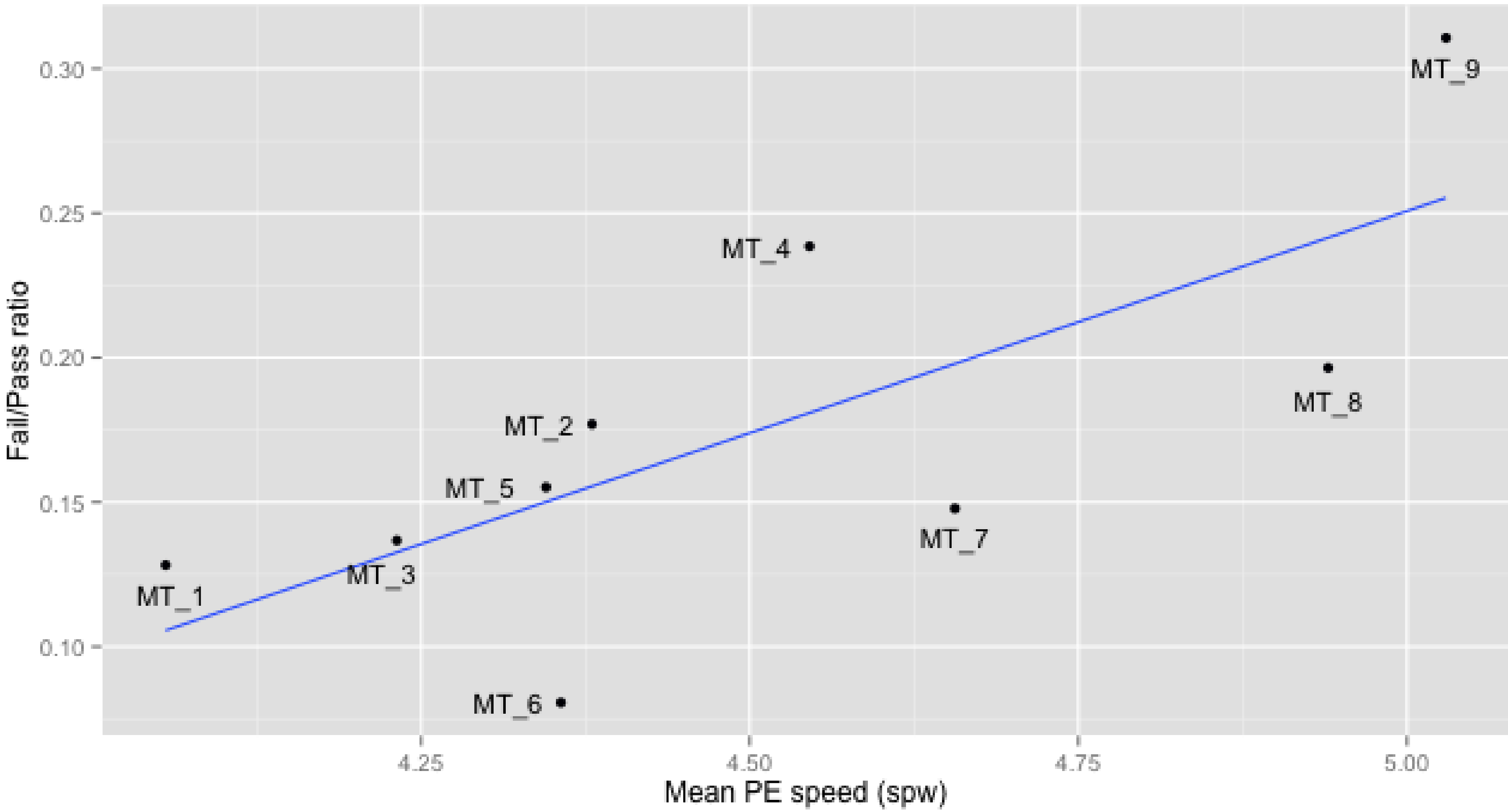
MT Quality and Productivity



BLEU against PE speed and regression line with 95% confidence bounds

+1 BLEU ↔ decrease in PE time of ∼0.16 sec/word, or 3-4% speed-up

MT Quality and PE Quality



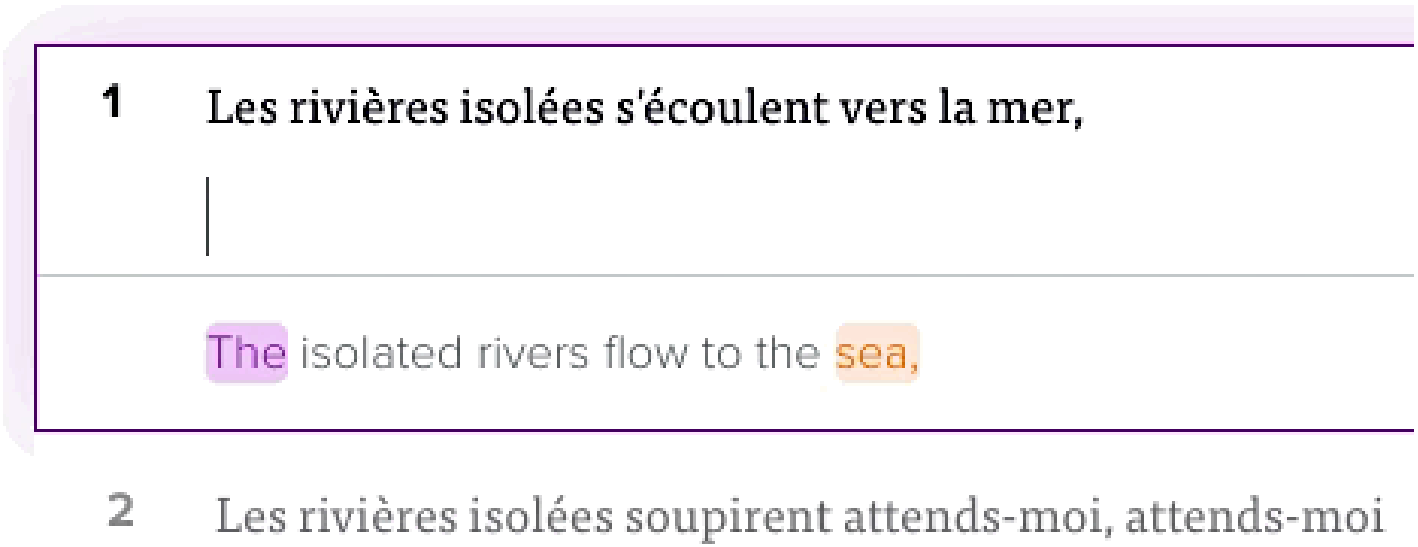
better MT $\leftrightarrow$ fewer post-editing errors

Translator Variability

	HTER	Edit Rate	PE speed (spw)	MQM Score	Fail	Pass
TR1	44.79	2.29	4.57	98.65	10	124
TR2	42.76	3.33	4.14	97.13	23	102
TR3	34.18	2.05	3.25	96.50	26	106
TR4	49.90	3.52	2.98	98.10	17	120
TR5	54.28	4.72	4.68	97.45	17	119
TR6	37.14	2.78	**2.86**	97.43	24	113
TR7	39.18	2.23	**6.36**	97.92	18	112
TR8	50.77	7.63	6.29	97.20	19	117
TR9	39.21	2.81	5.45	96.48	22	113

- Higher variability between translators than between MT systems

Interactive Translation Prediction



GIF animation: https://lilt.com/translators

Interactive Translation Prediction

- Interactive typing of the translation with next-word/phrase prediction
- Even more essential to use adaptive and interactive MT
- Studies inconclusive whether this is an improvement in quality/productivity over MTPE
- Current projects
  - Casmacat
  - Lilt.com
- Earlier projects
  - TransType/TransType2
  - Caitra

Interactive Translation Prediction

- Closer to, but still not hybrid-intelligence translation (Dillinger)
  - Humans + machines working together
  - No hybridization - two separate intelligences
  - No strategic use of best of both
  - Real-time interaction
  - Human in the driver seat

INTERACTIVE TRANSLATION PREDICTION DEMO WITH LILT.COM

Interactivity

- Traditional professional translation approaches
  - translation from scratch
  - post-editing translation memory match
  - post-editing machine translation output

- More interactive collaboration between machine and professional?

Interactive Machine Translation

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

|

Interactive Machine Translation

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

| He

Interactive Machine Translation

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

He | has

Interactive Machine Translation

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

He has | for months

Interactive Machine Translation

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

He planned |

Interactive Machine Translation

Input Sentence

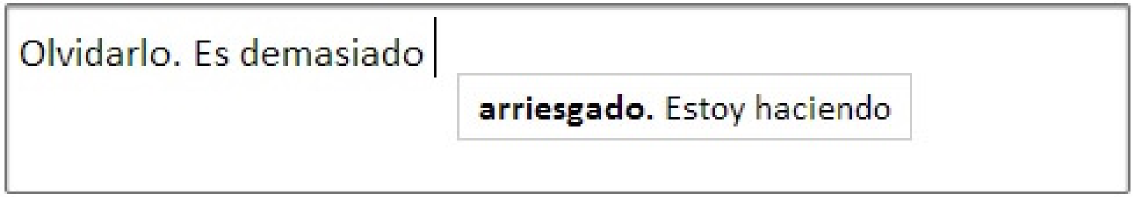
Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

He planned | for months

Visualization

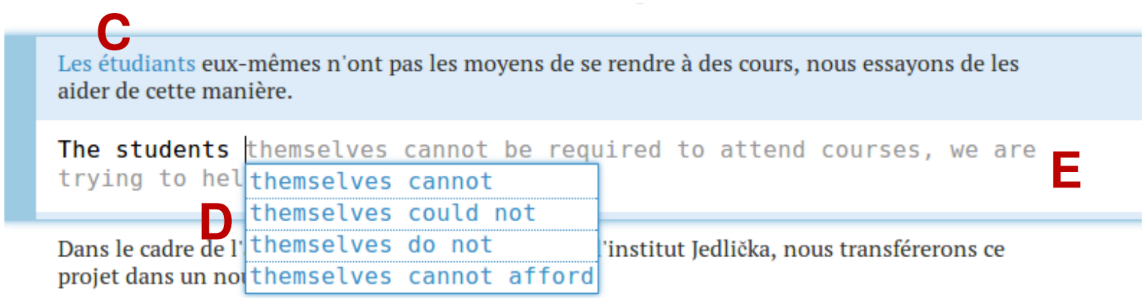
- Show n next words



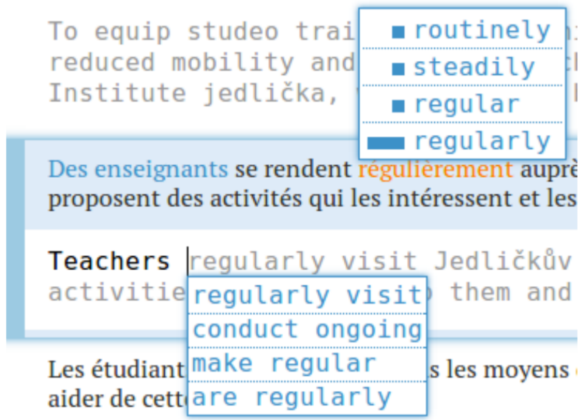
- Show rest of sentence

Spence Green's Lilt System

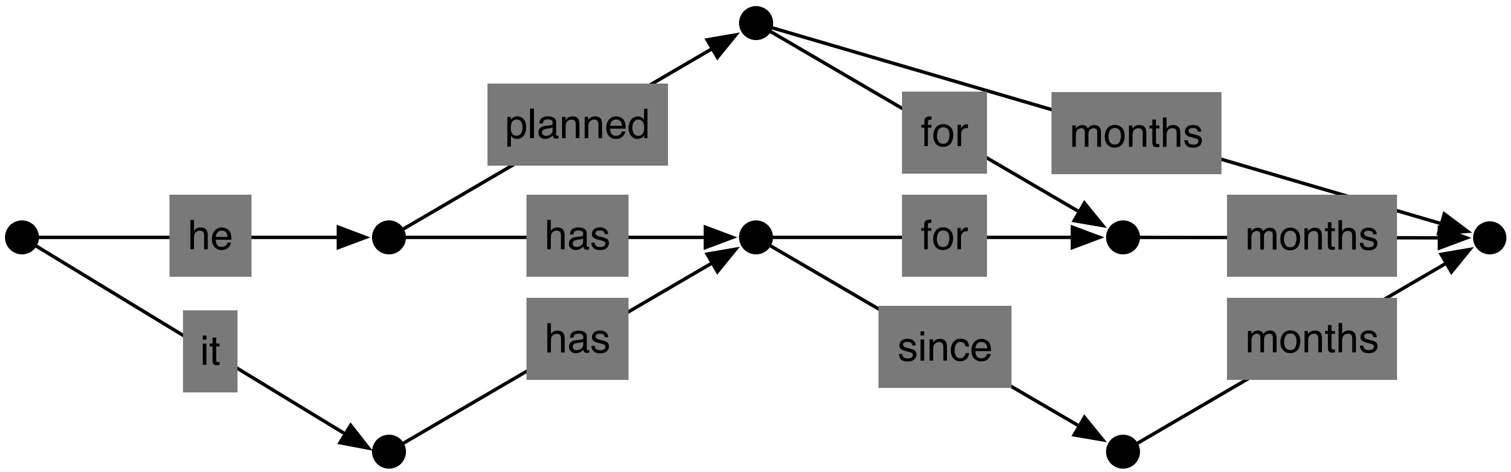
- Show alternate translation predictions



- Show alternate translations predictions with probabilities

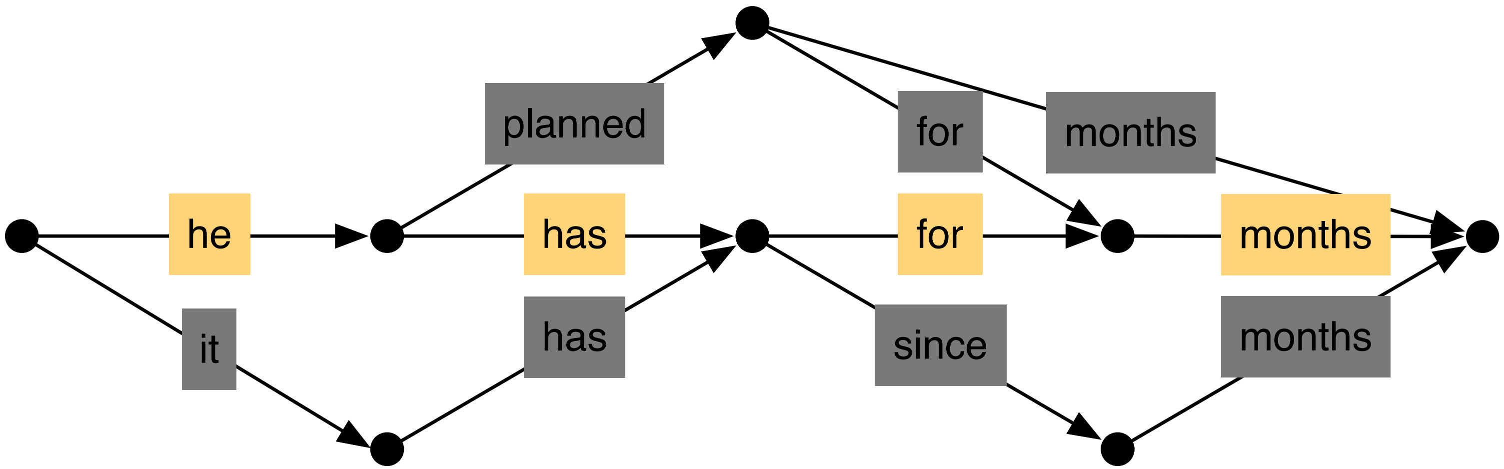


Prediction from Search Graph



Search for best translation creates a graph of possible translations

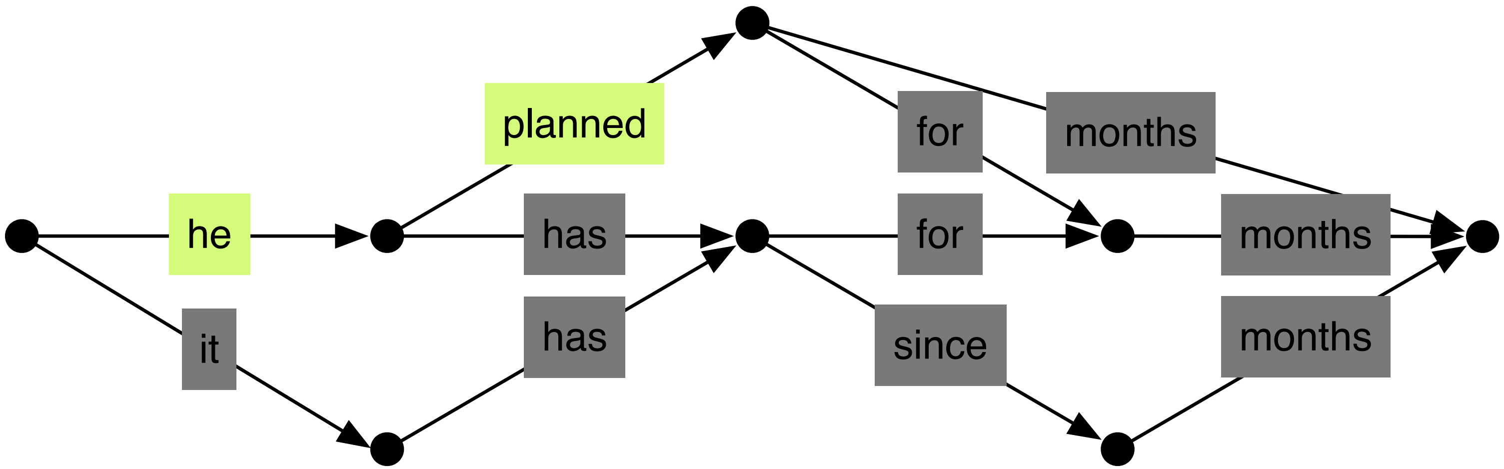
Prediction from Search Graph



One path in the graph is the best (according to the model)

This path is suggested to the user

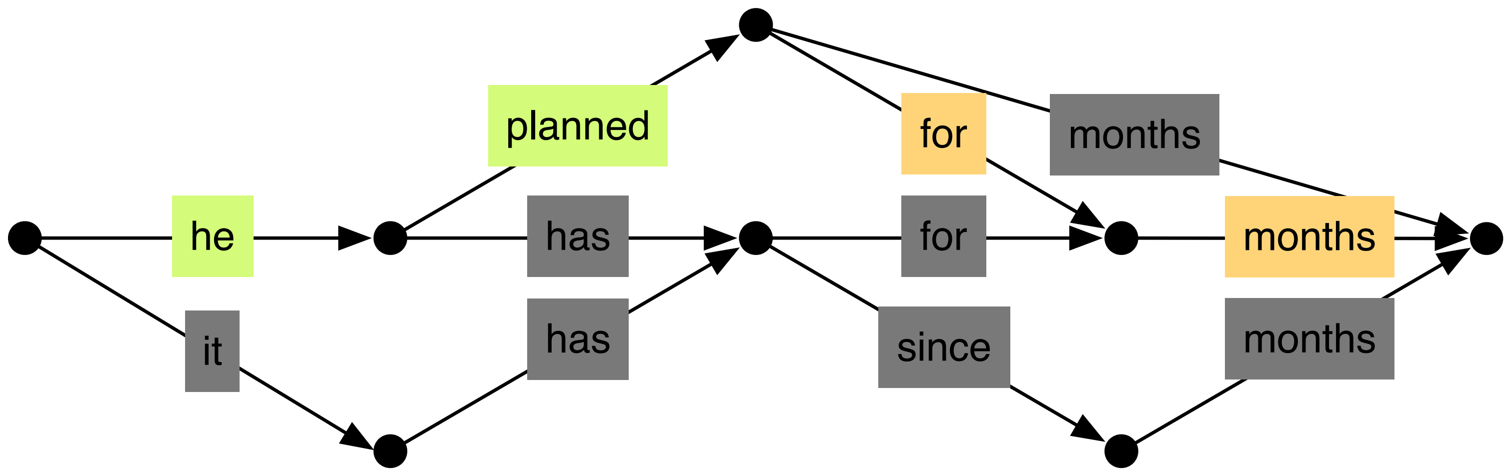
Prediction from Search Graph



The user may enter a different translation for the first words

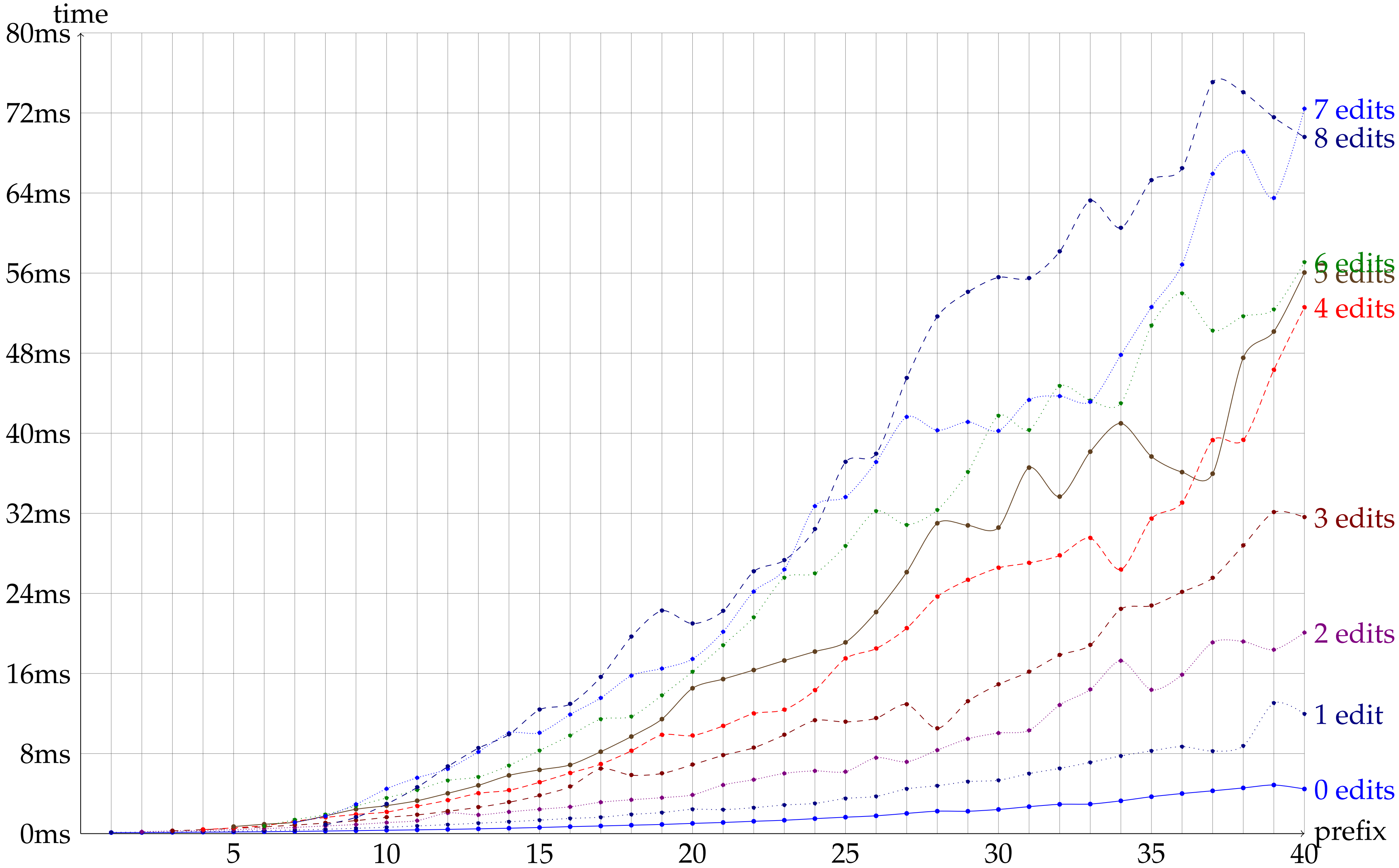
We have to find it in the graph

Prediction from Search Graph



We can predict the optimal completion (according to the model)

Run Time



time
80ms
72ms
64ms
56ms
48ms
40ms
32ms
24ms
16ms
8ms
0ms
5
10
15
20
25
30
35
40
prefix
7 edits
8 edits
6 edits
5 edits
4 edits
3 edits
2 edits
1 edit
0 edits

Word Alignment Visualization

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

He planned for months to give a lecture in Baltimore | in

Word Alignment Visualization

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten.

Professional Translator

He planned for months to give a lecture in Baltimore | in

Shading off Translated Material

Input Sentence

Er hat seit Monaten geplant, im November einen Vortrag in Baltimore zu halten .

Professional Translator

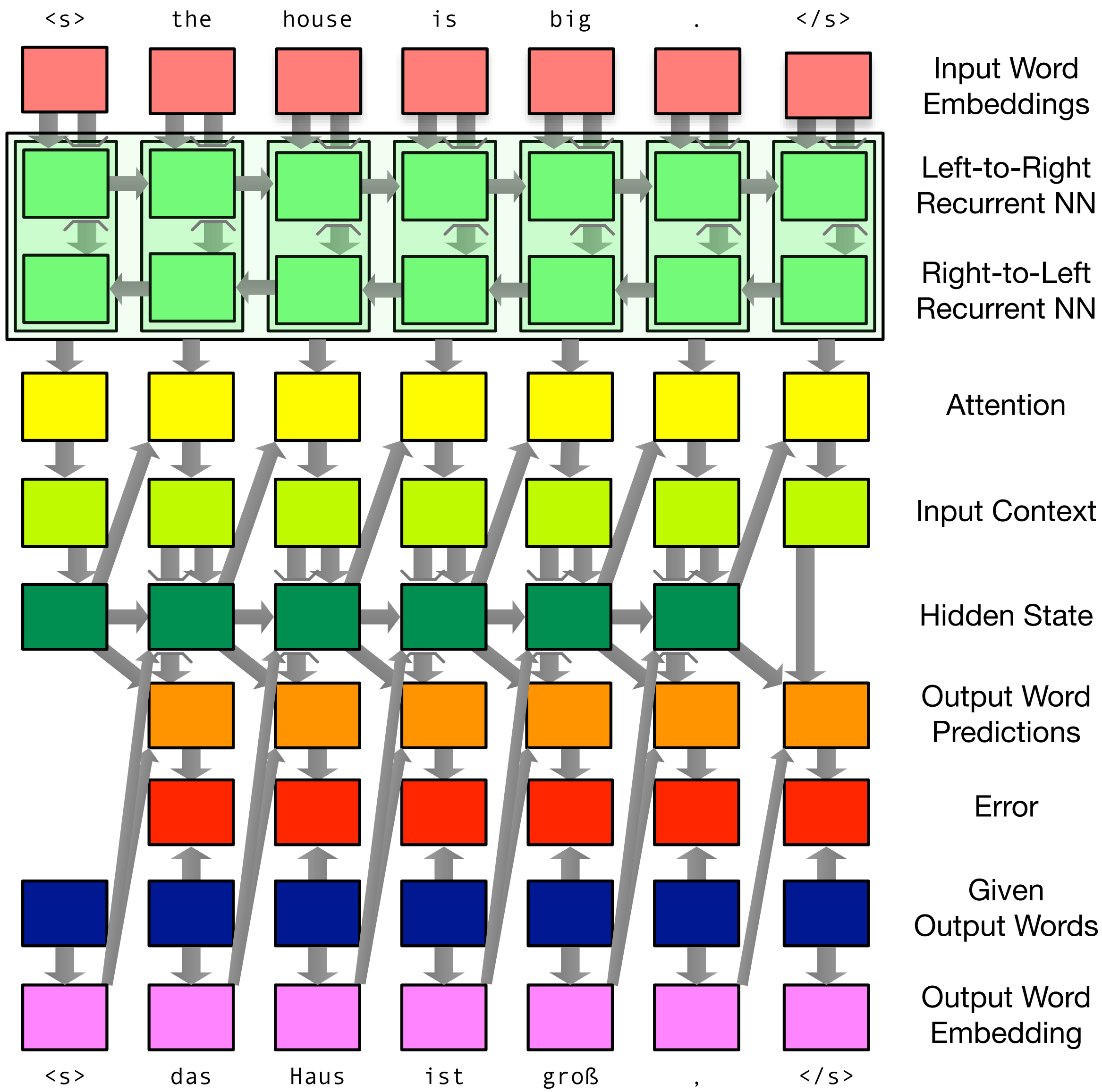
He planned for months to give a lecture in Baltimore | in

Some Observations

- How can we do this?
  - word alignments by-product of matching against search braph
  - automatic word alignments (as used in training)

- User feedback
  - users like interactive machine translation
  - ... but they may be slower than with post-editing machine translation
  - user like mouse-over word alignment highlighting
  - user do not like at-cursor word alignment highlighting

Neural Interactive Translation Prediction



Neural MT: Sequential Prediction

- The model produces words in sequence

$$p(\text{output}_t | \{\text{output}_1, \cdots, \text{output}_{t-1}\}, \vec{\text{input}}) = g(\hat{\text{output}}_{t-1}, \text{context}_t, \text{hidden}_t)$$

- Translation prediction: feed in user prefix

Example

Input: *Das Unternehmen sagte, dass es in diesem Monat mit Bewerbungsgesprächen beginnen wird und die Mitarbeiterzahl von Oktober bis Dezember steigt.*

	Correct	**Prediction**	**Prediction probability distribution**
✓	the	the	**the (99.2%)**
✓	company	company	**company (90.9%)**, firm (7.6%)
✓	said	said	**said (98.9%)**
✓	it	it	**it (42.6%)**, this (14.0%), that (13.1%), job (2.0%), the (1.7%), ...
✓	will	will	**will (77.5%)**, is (4.5%), started (2.5%), 's (2.0%), starts (1.8%), ...
✓	start	start	**start (49.6%)**, begin (46.7%)
	inter@@	job	job (16.1%), application (6.1%), en@@ (5.2%), out (4.8%), ...
✘	viewing	state	state (32.4%), related (5.8%), **viewing (3.4%)**, min@@ (2.0%), ...
✘	applicants	talks	talks (61.6%), interviews (6.4%), discussions (6.2%), ...
✓	this	this	**this (88.1%)**, so (1.9%), later (1.8%), that (1.1%)
✓	month	month	**month (99.4%)**
✘	,	and	and (90.8%), **, (7.7%)**
✘	with	and	and (42.6%), increasing (24.5%), rising (6.3%), **with (5.1%)**, ...
✓	staff	staff	**staff (22.8%)**, the (19.5%), employees (6.3%), employee (5.0%), ...
✘	levels	numbers	numbers (69.0%), **levels (3.3%)**, increasing (3.2%), ...
✘	rising	increasing	increasing (40.1%), **rising (35.3%)**, climbing (4.4%), rise (3.4%), ...
✓	from	from	**from (97.4%)**
✓	October	October	**October (81.3%)**, Oc@@ (12.8%), oc@@ (2.9%), Oct (1.2%)
✘	through	to	to (73.2%), **through (15.6%)**, until (8.7%)
✓	December	December	**December (85.6%)**, Dec (8.0%), to (5.1%)
✓	.	.	**. (97.5%)**

Knowles and Koehn [AMTA 2016]

- Better prediction accuracy, even when systems have same BLEU score (state-of-the-art German-English systems, compared to search graph matching)

System	Configuration	BLEU	Word Prediction Accuracy	Letter Prediction Accuracy
Neural	no beam search	34.5	61.6%	86.8%
	beam size 12	36.2	63.6%	87.4%
Phrase-based	-	34.5	43.3%	72.8%

Logging functions

- Different types of events are saved in the logging.
  - configuration and statistics
  - start and stop session
  - segment opened and closed
  - text, key strokes, and mouse events
  - scroll and resize
  - search and replace
  - suggestions loaded and suggestion chosen
  - interactive translation prediction
  - gaze and fixation from eye tracker

Logging functions

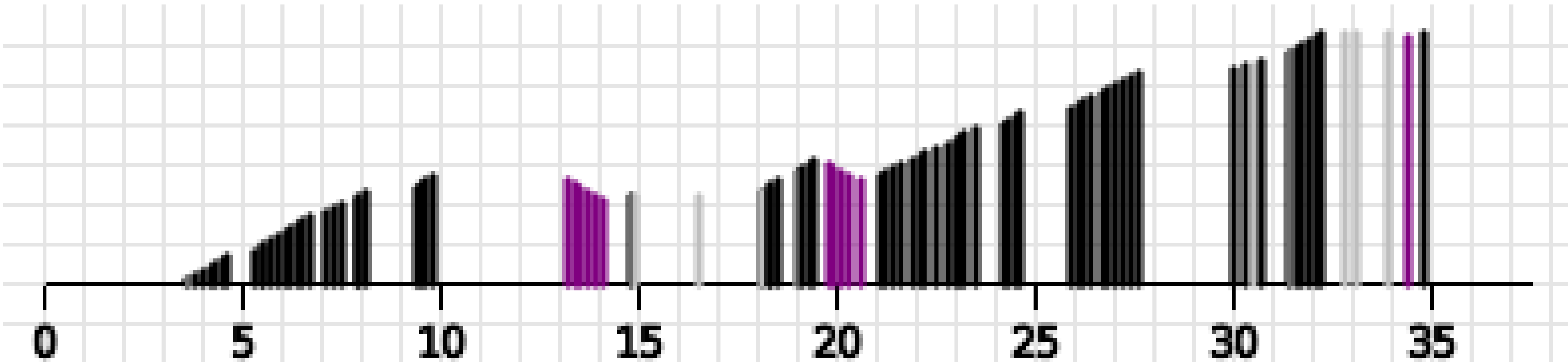
- In every event we save:
  - Type
  - In which element was produced
  - Time

- Special attributes are kept for some types of events
  - Diff of a text change
  - Current cursor position
  - Character looked at
  - Clicked UI element
  - Selected text

⇒ Full replay of user session is possible

Keystroke Log

Input: Au premier semestre, l'avionneur a livré 97 avions.
Output: The manufacturer has delivered 97 planes during the first half.



(37.5 sec, 3.4 sec/word)

black: keystroke, purple: deletion, grey: cursor move
height: length of sentence

Example of Quality Judgments

Src.	Sans se démonter, il s'est montré concis et précis.	
MT	Without dismantle, it has been concise and accurate.	
1/3	Without fail, he has been concise and accurate.	*(Prediction+Options, L2a)*
4/0	Without getting flustered, he showed himself to be concise and precise.	*(Unassisted, L2b)*
4/0	Without falling apart, he has shown himself to be concise and accurate.	*(Postedit, L2c)*
1/3	Unswayable, he has shown himself to be concise and to the point.	*(Options, L2d)*
0/4	Without showing off, he showed himself to be concise and precise.	*(Prediction, L2e)*
1/3	Without dismantling himself, he presented himself consistent and precise.	*(Prediction+Options, L1a)*
2/2	He showed himself concise and precise.	*(Unassisted, L1b)*
3/1	Nothing daunted, he has been concise and accurate.	*(Postedit, L1c)*
3/1	Without losing face, he remained focused and specific.	*(Options, L1d)*
3/1	Without becoming flustered, he showed himself concise and precise.	*(Prediction, L1e)*

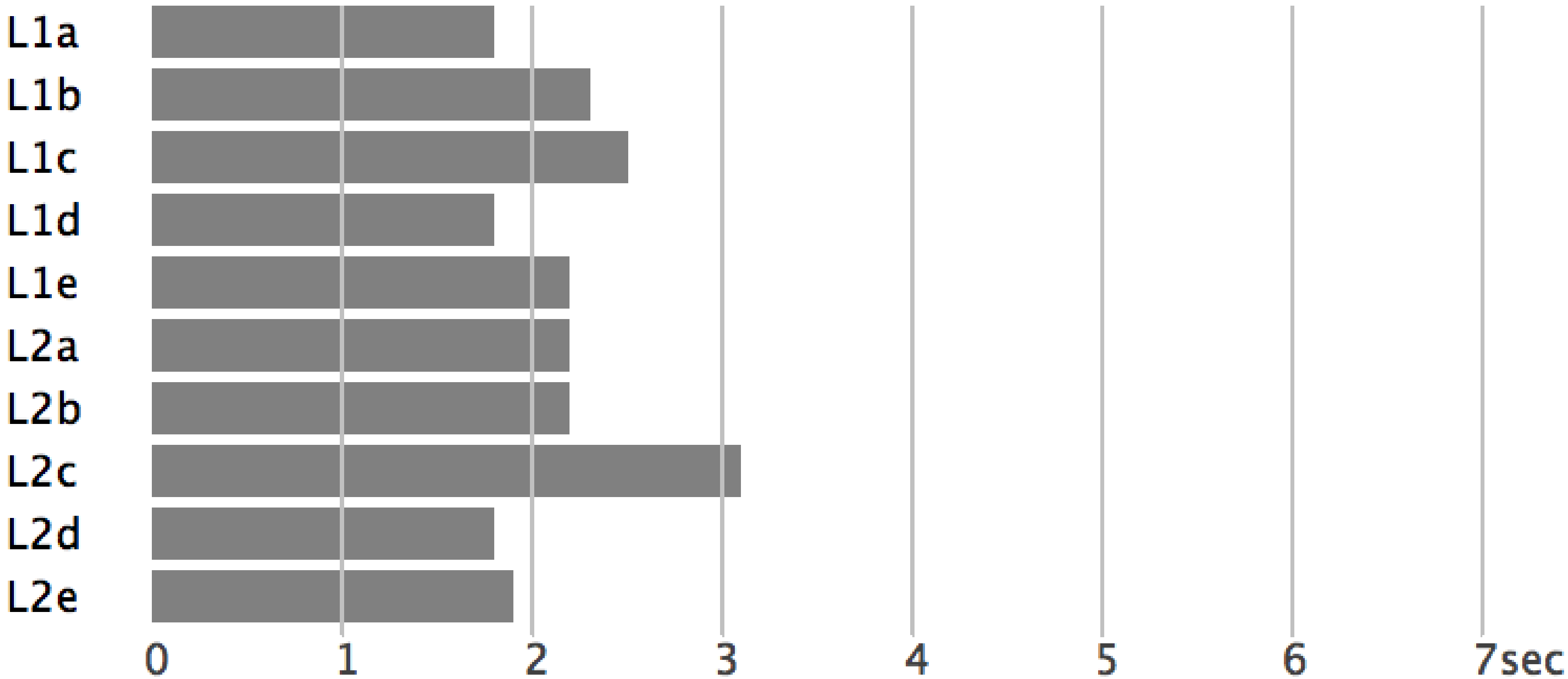
Main Measure: Productivity

Assistance	Speed	Quality
Unassisted	4.4s/word	47% correct
Postedit	2.7s (-1.7s)	55% (+8%)
Options	3.7s (-0.7s)	51% (+4%)
Prediction	3.2s (-1.2s)	54% (+7%)
Prediction+Options	3.3s (-1.1s)	53% (+6%)

Faster and Better, Mostly

User	Unassisted	Postedit		Options		Prediction		Prediction+Options	
L1a	3.3sec/word 23% correct	**1.2s** **39%**	**-2.2s** **+16%)**	**2.3s** **45%**	**-1.0s** **+22%**	**1.1s** **30%**	**-2.2s** **+7%)**	**2.4s** **44%**	**-0.9s** **+21%**
L1b	7.7sec/word 35% correct	**4.5s** **48%**	**-3.2s)** **+13%**	**4.5s** **55%**	**-3.3s** **+20%**	**2.7s** **61%**	**-5.1s** **+26%**	**4.8s** **41%**	**-3.0s** **+6%**
L1c	3.9sec/word 50% correct	**1.9s** **61%**	**-2.0s** **+11%**	**3.8s** **54%**	**-0.1s** **+4%**	**3.1s** **64%**	**-0.8s** **+14%**	**2.5s** **61%**	**-1.4s** **+11%**
L1d	2.8sec/word 38% correct	**2.0s** **46%**	**-0.7s** **+8%**	2.9s 59%	(+0.1s) (+21%)	2.4s 37%	(-0.4s) (-1%)	**1.8s** **45%**	**-1.0s** **+7%**
L1e	5.2sec/word 58% correct	**3.9s** **64%**	**-1.3s** **+6%**	4.9s 56%	(-0.2s) (-2%)	**3.5s** **62%**	**-1.7s** **+4%**	4.6s 56%	(-0.5s) (-2%)
L2a	5.7sec/word 16% correct	**1.8s** **50%**	**-3.9s** **+34%**	**2.5s** **34%**	**-3.2s** **+18%**	**2.7s** **40%**	**-3.0s** **+24%**	**2.8s** **50%**	**-2.9s** **+34%**
L2b	3.2sec/word 64% correct	2.8s 56%	(-0.4s) (-8%)	*3.5s* *60%*	*+0.3s* *-4%*	*6.0s* *61%*	*+2.8s* *-3%*	*4.6s* *57%*	*+1.4s* *-7%*
L2c	5.8sec/word 52% correct	**2.9s** **53%**	**-3.0s** **+1%**	4.6s 37%	(-1.2s) (-15%)	**4.1s** **59%**	**-1.7s** **+7%**	**2.7s** **53%**	**-3.1s** **+1%**
L2d	3.4sec/word 49% correct	3.1s 49%	(-0.3s) (+0%)	4.3s 51%	(+0.9s) (+2%)	3.8s 53%	(+0.4s) (+4%)	3.7s 58%	(+0.3s) (+9%)
L2e	2.8sec/word 68% correct	**2.6s** **79%**	**-0.2s** **+11%**	*3.5s* *59%*	*+0.7s* *-9%*	2.8s 64%	(-0.0s) (-4%)	*3.0s* *66%*	*+0.2s* *-2%*
avg.	4.4sec/word 47% correct	**2.7s** **55%**	**-1.7s** **+8%**	**3.7s** **51%**	**-0.7s** **+4%**	**3.2s** **54%**	**-1.2s** **+7%**	**3.3s** **53%**	**-1.1s** **+6%**

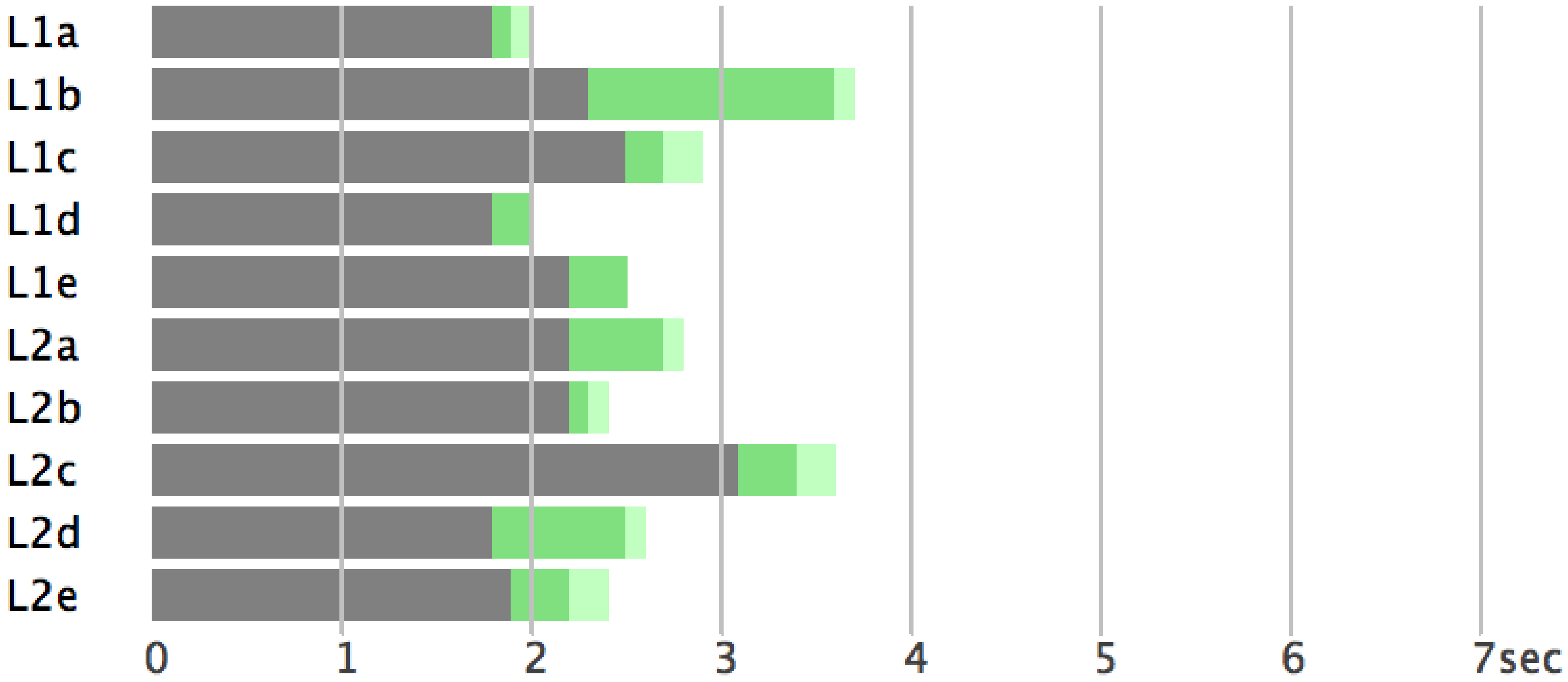
Unassisted Novice Translators



L1 = native French, L2 = native English, average time per input word

only typing

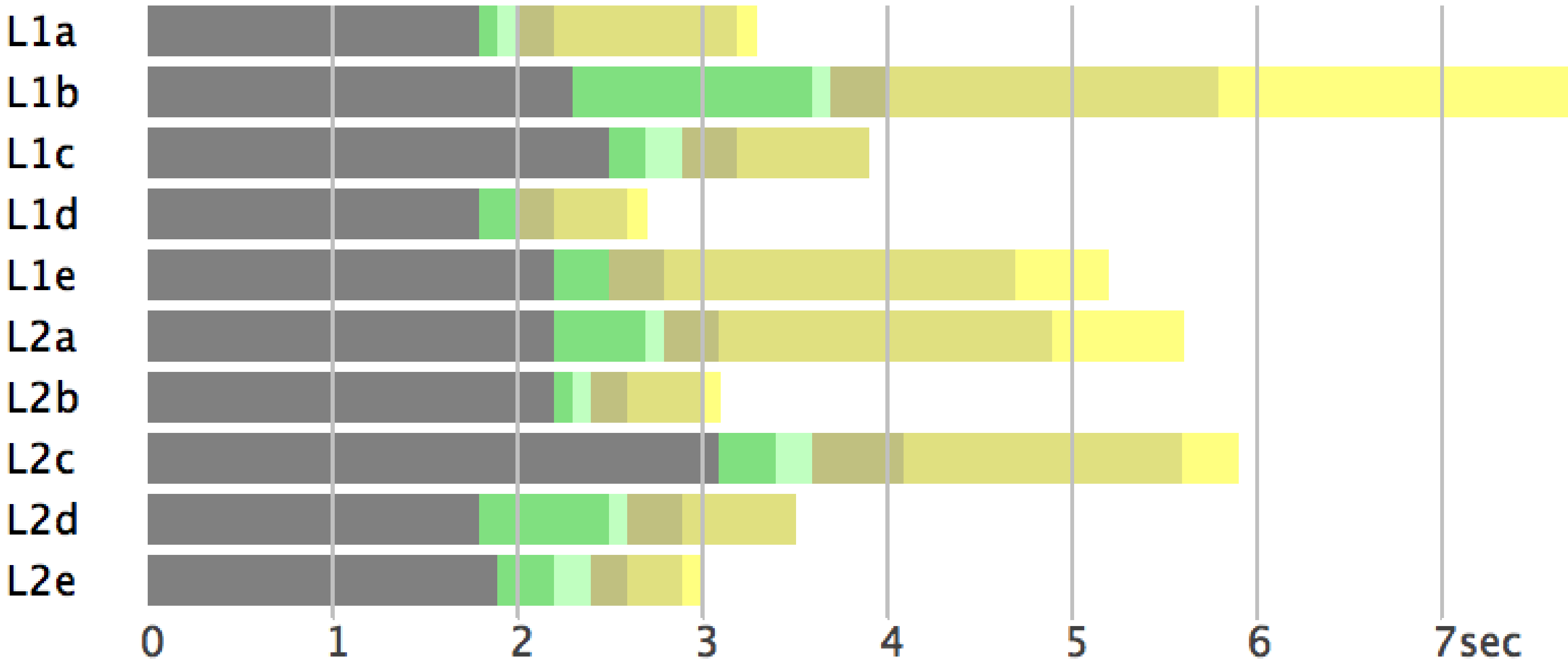
Unassisted Novice Translators



L1 = native French, L2 = native English, average time per input word

typing, initial and final pauses

Unassisted Novice Translators



L1 = native French, L2 = native English, average time per input word

typing, initial and final pauses, short, medium, and long pauses

most time difference on intermediate pauses

Activities: Native French User L1b

User: L1b	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab
Unassisted	7.7s	1.3s	0.1s	0.3s	1.8s	1.9s	2.3s	-	-
Postedit	4.5s	1.5s	0.4s	0.1s	1.0s	0.4s	1.1s	-	-
Options	4.5s	0.6s	0.1s	0.4s	0.9s	0.7s	1.5s	0.4s	-
Prediction	2.7s	0.3s	0.3s	0.2s	0.7s	0.1s	0.6s	-	0.4s
Prediction+Options	4.8s	0.6s	0.4s	0.4s	1.3s	0.5s	0.9s	0.5s	0.2s

Activities: Native French User L1b

User: L1b	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab
Unassisted	7.7s	1.3s	0.1s	0.3s	1.8s	1.9s	2.3s	-	-
Postedit	4.5s	1.5s	0.4s	0.1s	1.0s	0.4s	1.1s	-	-
Options	4.5s	0.6s	0.1s	0.4s	0.9s	0.7s	1.5s	0.4s	-
Prediction	2.7s	0.3s	0.3s	0.2s	0.7s	0.1s	0.6s	-	0.4s
Prediction+Options	4.8s	0.6s	0.4s	0.4s	1.3s	0.5s	0.9s	0.5s	0.2s

Slightly less time spent on typing

Activities: Native French User L1b

User: L1b	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab
Unassisted	7.7s	1.3s	0.1s	0.3s	1.8s	1.9s	2.3s	-	-
Postedit	4.5s	1.5s	0.4s	0.1s	1.0s	0.4s	1.1s	-	-
Options	4.5s	0.6s	0.1s	0.4s	0.9s	0.7s	1.5s	0.4s	-
Prediction	2.7s	0.3s	0.3s	0.2s	0.7s	0.1s	0.6s	-	0.4s
Prediction+Options	4.8s	0.6s	0.4s	0.4s	1.3s	0.5s	0.9s	0.5s	0.2s

Less pausing

Slightly less time spent on typing

Activities: Native French User L1b

User: L1b	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab
Unassisted	7.7s	1.3s	0.1s	0.3s	1.8s	1.9s	2.3s	-	-
Postedit	4.5s	1.5s	0.4s	0.1s	1.0s	0.4s	1.1s	-	-
Options	4.5s	0.6s	0.1s	0.4s	0.9s	0.7s	1.5s	0.4s	-
Prediction	2.7s	0.3s	0.3s	0.2s	0.7s	0.1s	0.6s	-	0.4s
Prediction+Options	4.8s	0.6s	0.4s	0.4s	1.3s	0.5s	0.9s	0.5s	0.2s

Less pausing

Especially less time in big pauses

Slightly less time spent on typing

References

- Andy Way. 2018. Quality Expectations of Machine Translation . In S. Castilho, J. Moorkens, F. Gaspari and S. Doherty (eds) *Translation Quality Assessment: From Principles to Practice*, Springer.
- TAUS/DCU Machine Translation Post-Editing Guidelines
- Mike Dillinger, Hybrid-intelligence translation: forging a real partnership between humans and machines, memoQFest 2017 keynote
- Mike Dillinger, "MT escaped from the Lab!", Keynote AMTA 2016
- Rebecca Knowles and Philipp Koehn, Neural Interactive Translation Prediction, Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), 2016.

References

- Spence Green, Jeffrey Heer and Christopher D. Manning, Natural Language Translation at the Intersection of AI and HCI, ACM Queue, HCI Vol. 13 No. 6 – June 2015
- Germán Sanchis-Trilles, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García- Martínez, Jesús González-Rubio, Robin L. Hill, Philipp Koehn, Luis A. Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Herve Saint-Amand, Chara Tsoukala. Interactive Translation Prediction vs. Conventional Post-editing in Practice: A Study with the CasMaCat Workbench. Machine Translation, 28(3-4):217-235, 2014