Preparing the KantanMT Community for migration to NeuralMT

OpenNMT Workshop Paris, March 2018

Tony O'Dowd
Chief **Architect**
KantanMT.com

KantanMT.com

No Hardware. No Software. No Hassle MT.

The KantanMT Community Migration to NeuralMT

Three Step Process

- **Step 1**
  - Release KantanNeural
- **Step 2**
  - Prove OpenNMT is better than MOSES SMT
- **Step 3**
  - Measure the Impact
- **Step 4**
  - Celebrate (purely optional)

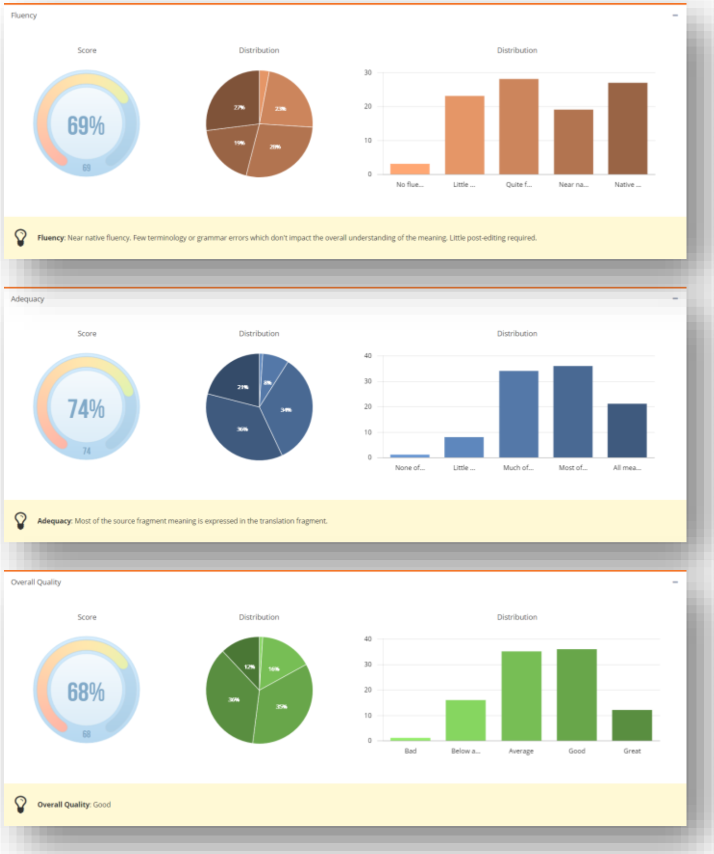
KantanMT.com

No Hardware. No Software. No Hassle MT.

Step 1 : Introduce KantanNeural™

Step 1: Release KantanNeural

- **KantanNeural**
  - Prototype I built Nov 2016
    - Nematus, ADAM, BPE, SCN
    - 12 Language Pairs released
  - Migrated to OpenNMT Jan 2017
    - Launched KantanNeural Mar 2017
    - 39 Language Pairs
- **KantanBuildAnalytics**
  - Remodelled to support Perplexity scores – June 2017
- **KantanLQR**
  - Extended to support Ranking and enhanced MQM factors
- **Other milestones**
  - Terminology support (Level 1) – April 2017
  - Tag/Placeholder support – May 2017
  - NMT Adaptation – Aug 2017
  - Super fast training time – Sept 2017



Fluency: Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little post-editing required.

Adequacy: Most of the source fragment meaning is expressed in the translation fragment.

Overall Quality: Good

KantanMT.com

No Hardware. No Software. No Hassle MT.

Step 2 : Prove the Hype!

Step 2: NMT is better than SMT?

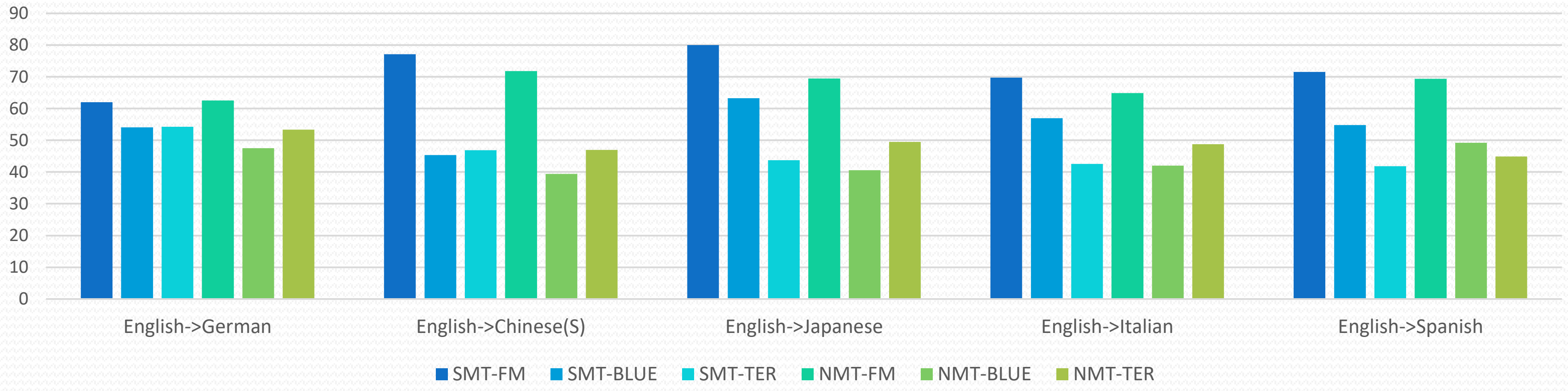
- **Experiment Setup**
  - Identical Training Data Sets
  - Identical Test Reference Sets
  - Automated Scores Used: F-Measure, TER, BLEU
  - Native Speaking, Professional Reviewers
  - NMT – KantanNeural™ – GPU Processors
  - SMT – KantanMT – CPU Processors
  - Translation Evaluation – KantanLQR™

Training Corpora

Language Arc	Parallel Sentences	TWC	UWC	Domain(s)
English->German	8,820,562	110,150,238	859,167	Legal/Medical
English->Chinese(Simplified)	6,522,064	84,426,931	956,864	Legal/Technical
English->Japanese	8,545,366	87,252,129	676,244	Legal/Technical
English->Italian	2,756,185	35,295,535	765,930	Medical
English->Spanish	3,681,332	44,917,538	952,089	Legal

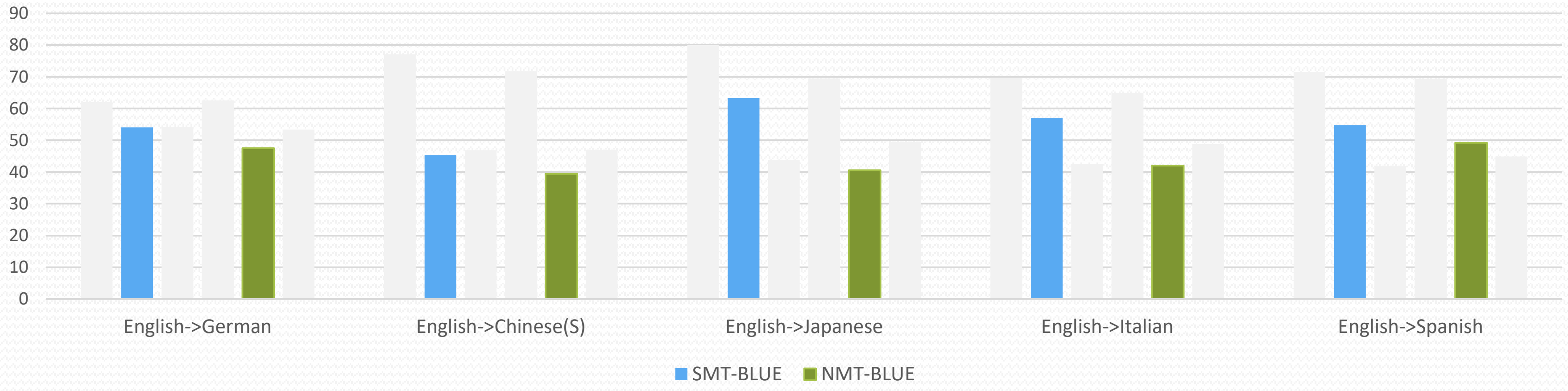
Training : Automated Scores

	SMT				NMT				
Language Arc	F-Measure	BLUE	TER	Time	F-Measure	BLUE	TER	Perplexity	Time
English->German	62.00	54.08	54.31	18	62.53	47.53	53.41	3.02	92
English->Chinese(Simplified)	77.16	45.36	46.85	6	71.85	39.39	47.01	2.00	10
English->Japanese	80.04	63.27	43.77	9	69.51	40.55	49.46	1.89	68
English->Italian	69.74	56.98	42.54	8	64.88	42.00	48.73	2.70	83
English->Spanish	71.53	54.78	41.87	9	69.41	49.24	44.89	2.59	71

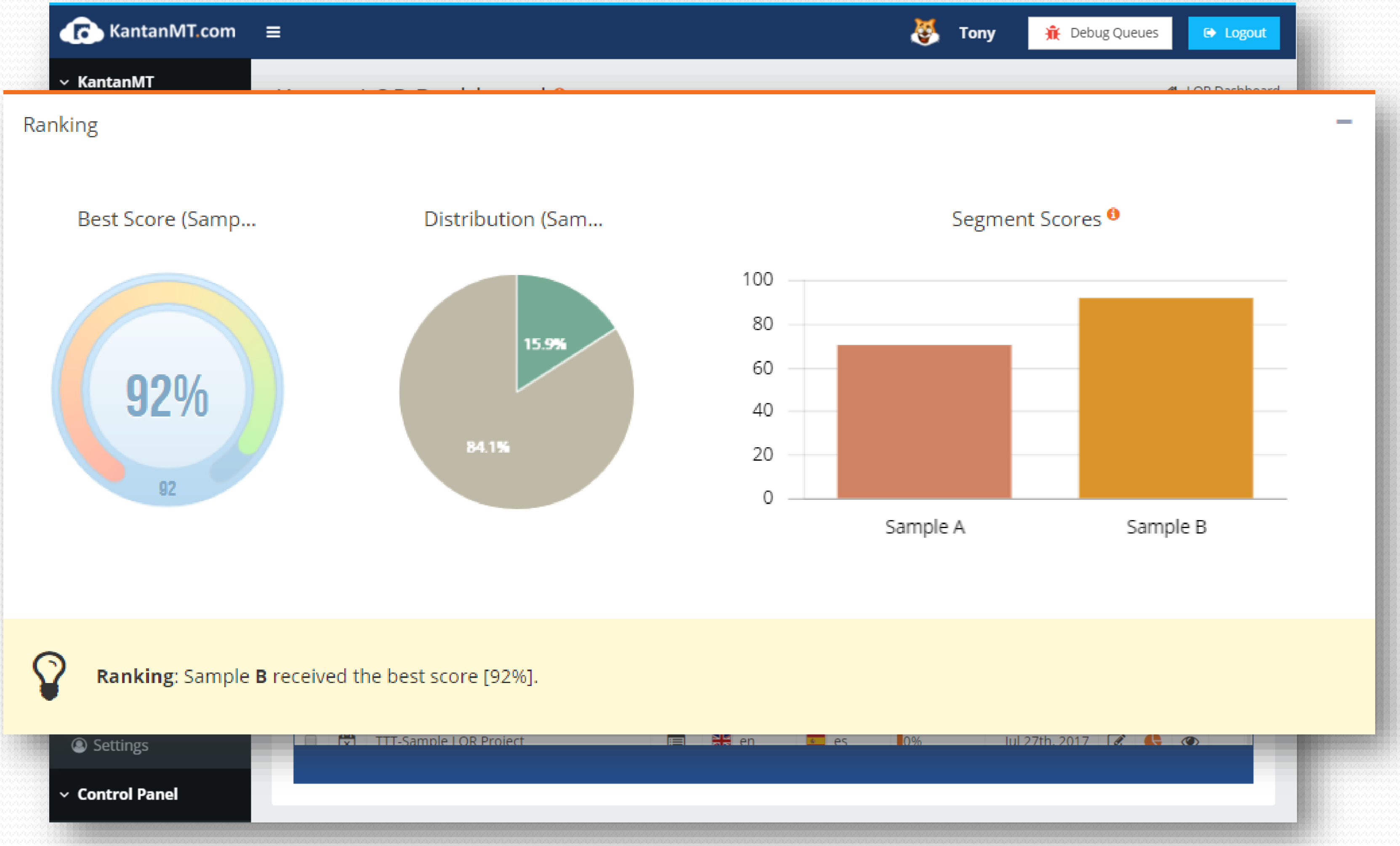


Training : Automated Scores

	SMT				NMT				
Language Arc	F-Measure	BLUE	TER	Time	F-Measure	BLUE	TER	Perplexity	Time
English->German	62.00	54.08	54.31	18	62.53	47.53	53.41	3.02	92
English->Chinese(Simplified)	77.16	45.36	46.85	6	71.85	39.39	47.01	2.00	10
English->Japanese	80.04	63.27	43.77	9	69.51	40.55	49.46	1.89	68
English->Italian	69.74	56.98	42.54	8	64.88	42.00	48.73	2.70	83
English->Spanish	71.53	54.78	41.87	9	69.41	49.24	44.89	2.59	71

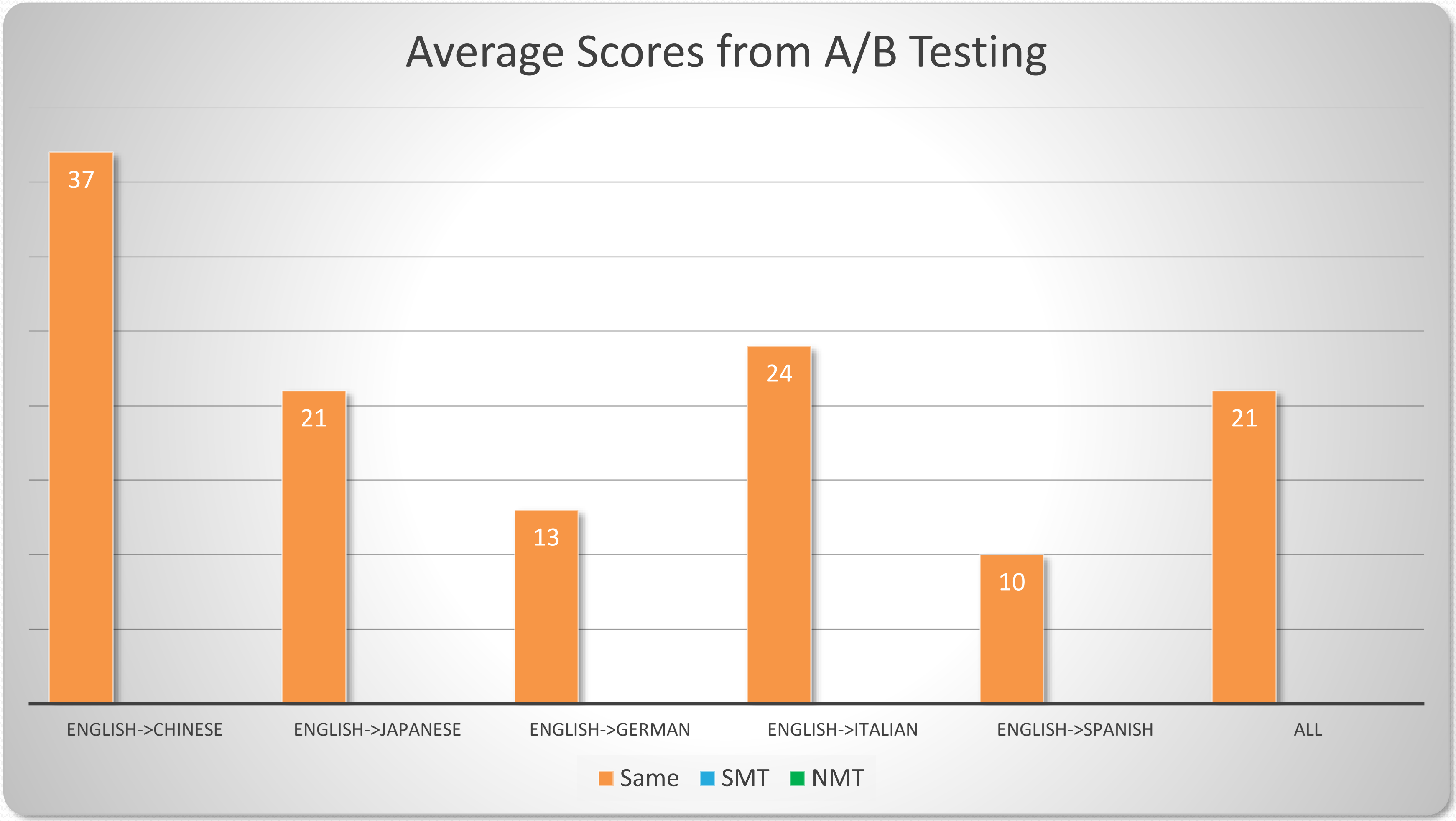


KantanLQR – 5 Teams of PHTs

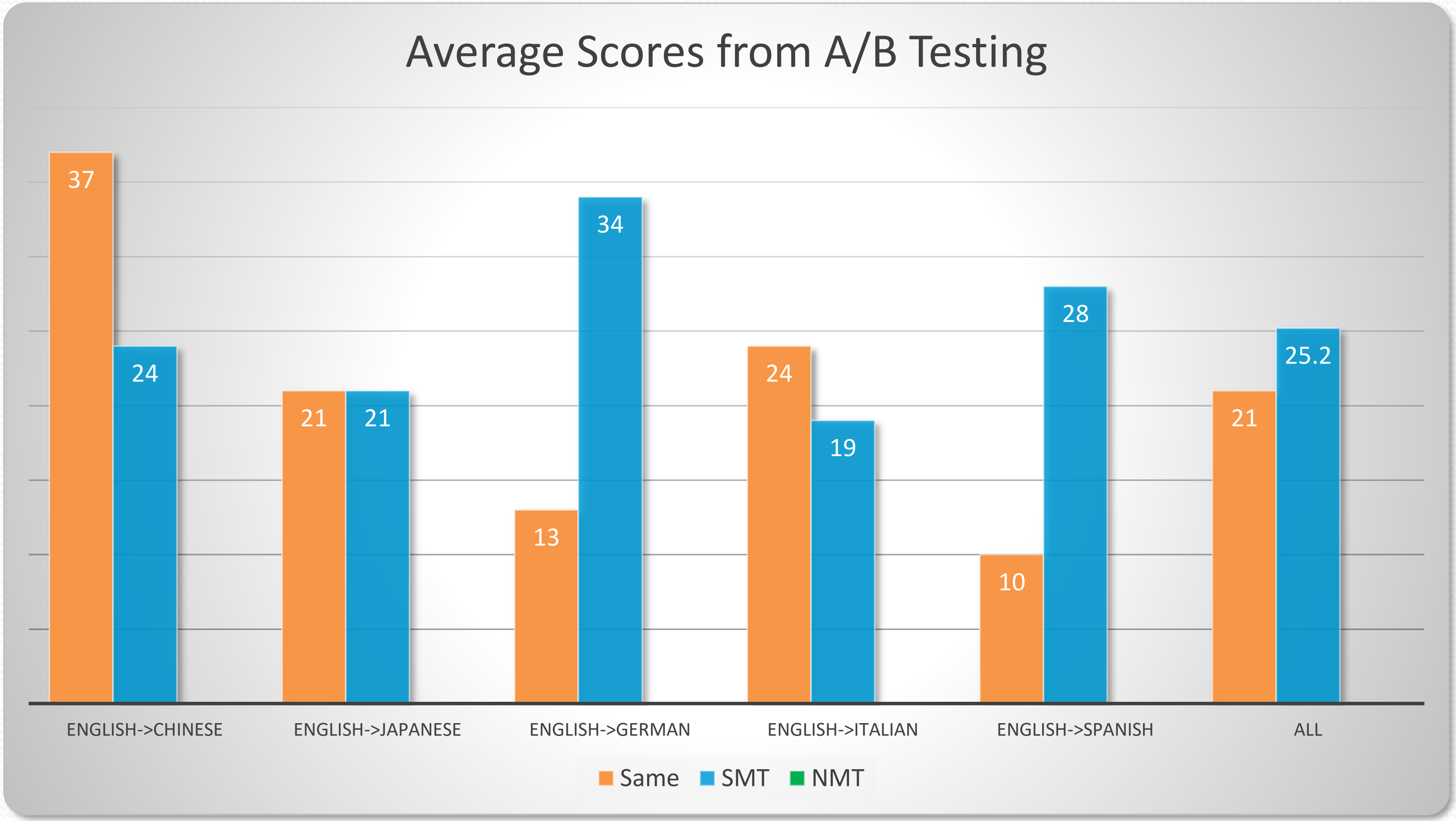


Ranking: Sample **B** received the best score [92%].

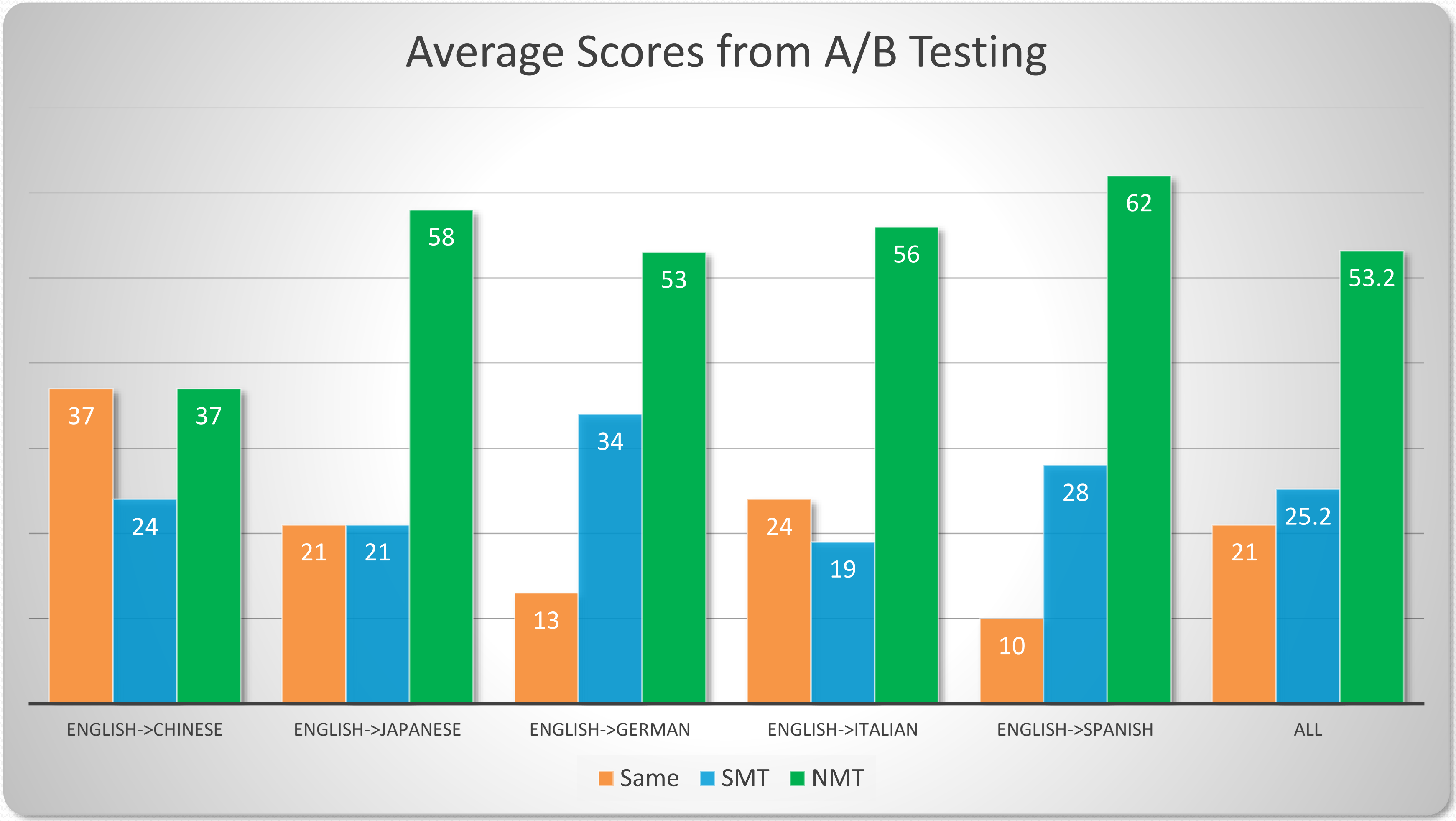
Ranking


Average Scores from A/B Testing
37
21
13
24
10
21
ENGLISH->CHINESE
ENGLISH->JAPANESE
ENGLISH->GERMAN
ENGLISH->ITALIAN
ENGLISH->SPANISH
ALL
Same
SMT
NMT

Ranking


Average Scores from A/B Testing
37
24
21
21
13
34
24
19
10
28
21
25.2
ENGLISH->CHINESE
ENGLISH->JAPANESE
ENGLISH->GERMAN
ENGLISH->ITALIAN
ENGLISH->SPANISH
ALL
Same
SMT
NMT

Ranking


Average Scores from A/B Testing
37
24
37
21
21
58
13
34
53
24
19
56
10
28
62
21
25.2
53.2
ENGLISH->CHINESE
ENGLISH->JAPANESE
ENGLISH->GERMAN
ENGLISH->ITALIAN
ENGLISH->SPANISH
ALL
Same
SMT
NMT

KantanMT.com

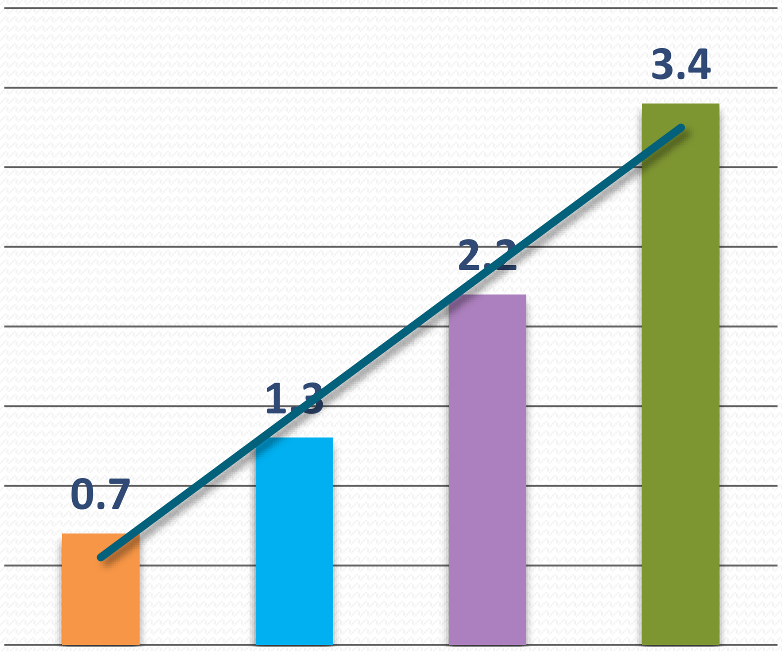
No Hardware. No Software. No Hassle MT.

Step 3 : Measure the Impact

Step 3: Measure the Impact

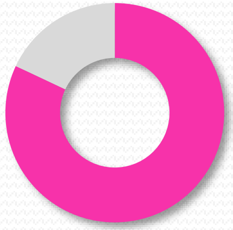
Words Translated

(in Billions)



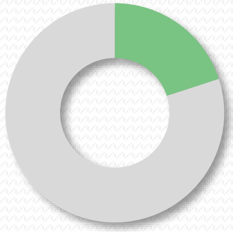
Engine Breakdown

(% percentage)



KantanMT™

MOSES Based SMT

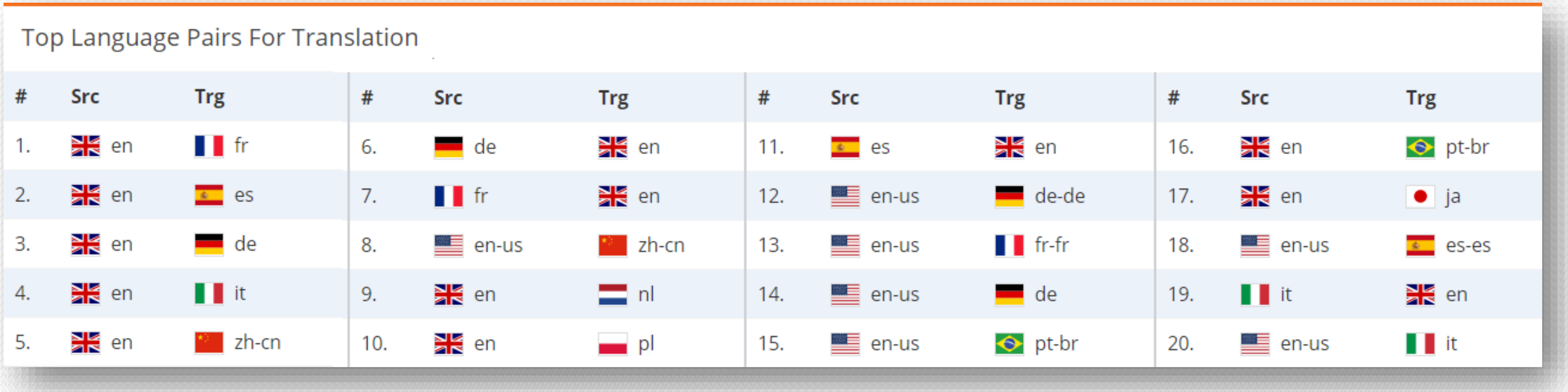


KantanNeural™

OpenNMT based NMT

Step 3: Measure the Impact

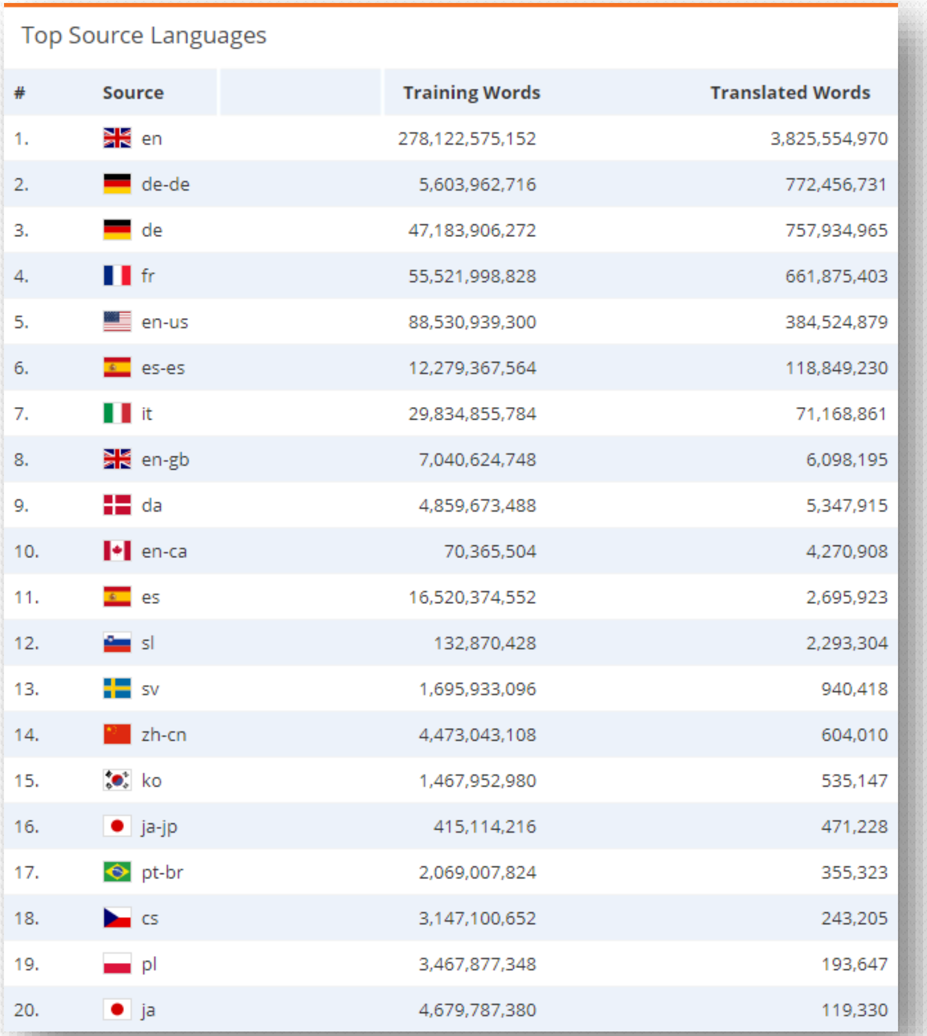
- **Top 20 NMT Language Arcs**



Top Language Pairs For Translation

#	Src	Trg	#	Src	Trg	#	Src	Trg	#	Src	Trg
1.	en	fr	6.	de	en	11.	es	en	16.	en	pt-br
2.	en	es	7.	fr	en	12.	en-us	de-de	17.	en	ja
3.	en	de	8.	en-us	zh-cn	13.	en-us	fr-fr	18.	en-us	es-es
4.	en	it	9.	en	nl	14.	en-us	de	19.	it	en
5.	en	zh-cn	10.	en	pl	15.	en-us	pt-br	20.	en-us	it

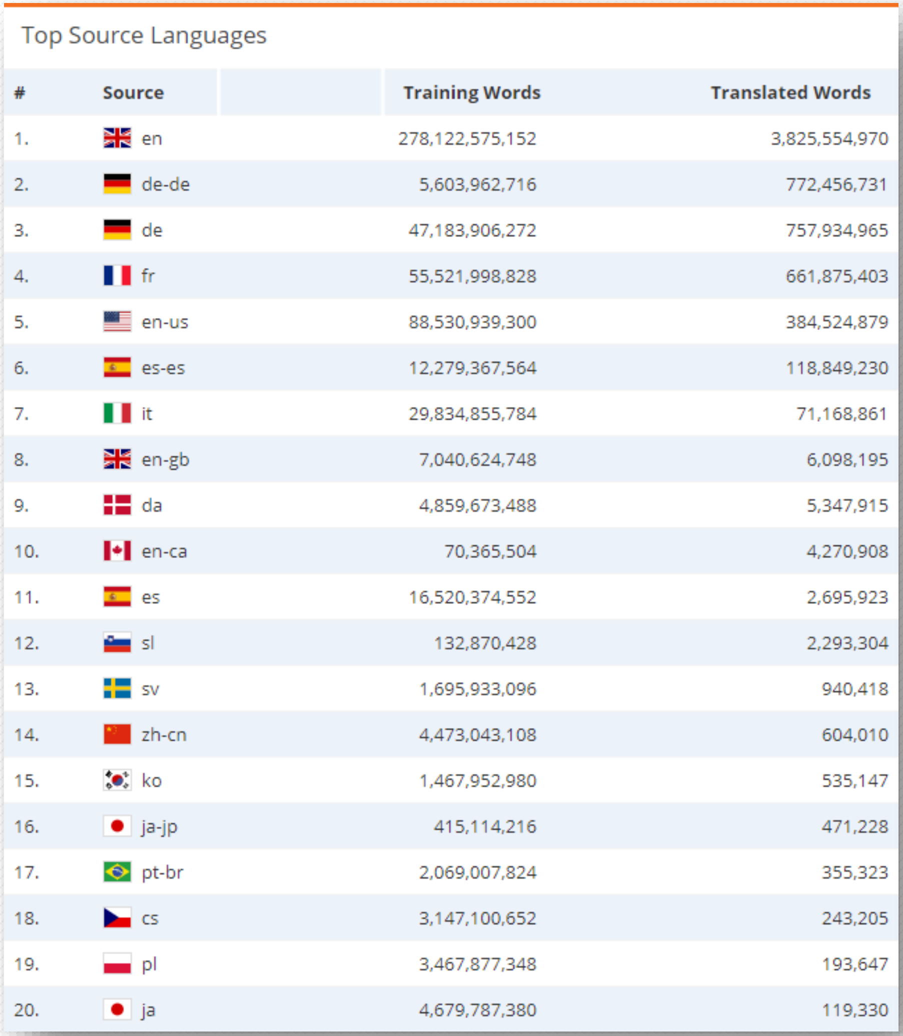
Step 3: Measure the Impact



Top Source Languages

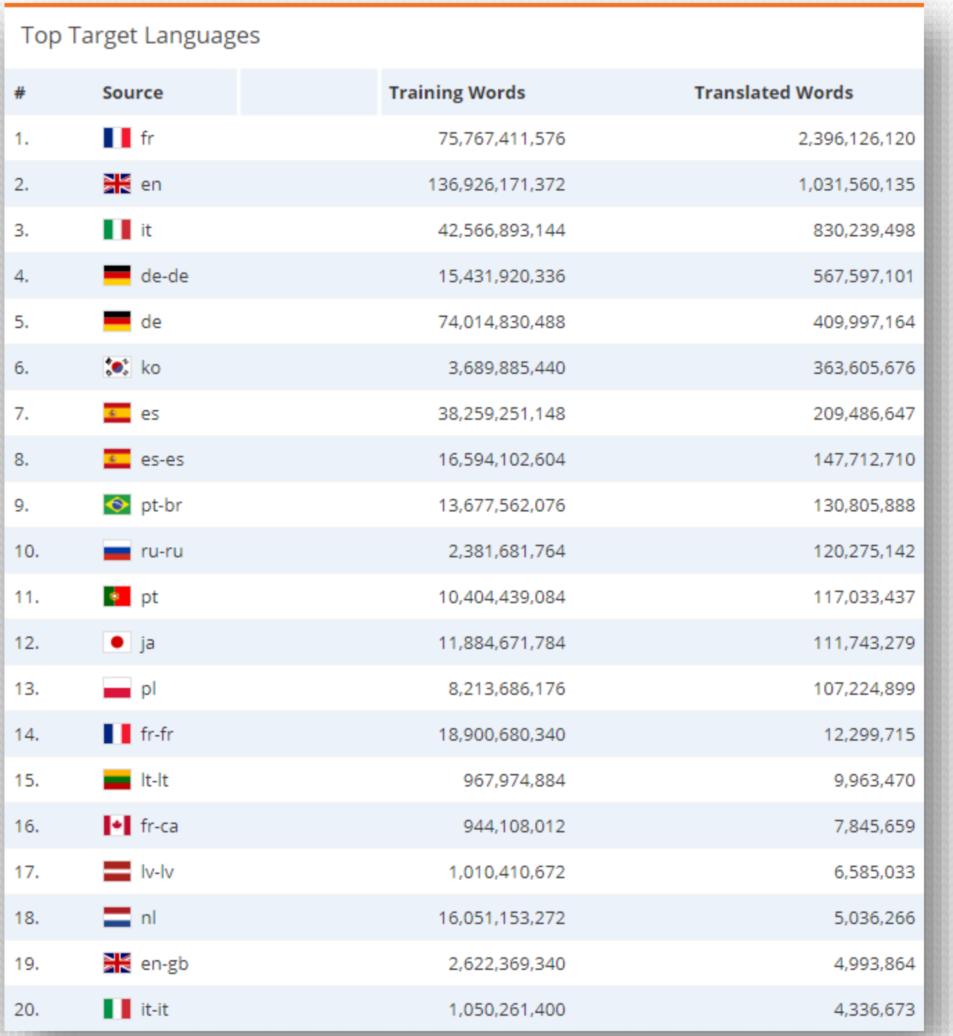
#	Source		Training Words	Translated Words
1.	en		278,122,575,152	3,825,554,970
2.	de-de		5,603,962,716	772,456,731
3.	de		47,183,906,272	757,934,965
4.	fr		55,521,998,828	661,875,403
5.	en-us		88,530,939,300	384,524,879
6.	es-es		12,279,367,564	118,849,230
7.	it		29,834,855,784	71,168,861
8.	en-gb		7,040,624,748	6,098,195
9.	da		4,859,673,488	5,347,915
10.	en-ca		70,365,504	4,270,908
11.	es		16,520,374,552	2,695,923
12.	sl		132,870,428	2,293,304
13.	sv		1,695,933,096	940,418
14.	zh-cn		4,473,043,108	604,010
15.	ko		1,467,952,980	535,147
16.	ja-jp		415,114,216	471,228
17.	pt-br		2,069,007,824	355,323
18.	cs		3,147,100,652	243,205
19.	pl		3,467,877,348	193,647
20.	ja		4,679,787,380	119,330

Step 3: Measure the Impact



Top Source Languages

#	Source	Training Words	Translated Words
1.	en	278,122,575,152	3,825,554,970
2.	de-de	5,603,962,716	772,456,731
3.	de	47,183,906,272	757,934,965
4.	fr	55,521,998,828	661,875,403
5.	en-us	88,530,939,300	384,524,879
6.	es-es	12,279,367,564	118,849,230
7.	it	29,834,855,784	71,168,861
8.	en-gb	7,040,624,748	6,098,195
9.	da	4,859,673,488	5,347,915
10.	en-ca	70,365,504	4,270,908
11.	es	16,520,374,552	2,695,923
12.	sl	132,870,428	2,293,304
13.	sv	1,695,933,096	940,418
14.	zh-cn	4,473,043,108	604,010
15.	ko	1,467,952,980	535,147
16.	ja-jp	415,114,216	471,228
17.	pt-br	2,069,007,824	355,323
18.	cs	3,147,100,652	243,205
19.	pl	3,467,877,348	193,647
20.	ja	4,679,787,380	119,330



Top Target Languages

#	Source	Training Words	Translated Words
1.	fr	75,767,411,576	2,396,126,120
2.	en	136,926,171,372	1,031,560,135
3.	it	42,566,893,144	830,239,498
4.	de-de	15,431,920,336	567,597,101
5.	de	74,014,830,488	409,997,164
6.	ko	3,689,885,440	363,605,676
7.	es	38,259,251,148	209,486,647
8.	es-es	16,594,102,604	147,712,710
9.	pt-br	13,677,562,076	130,805,888
10.	ru-ru	2,381,681,764	120,275,142
11.	pt	10,404,439,084	117,033,437
12.	ja	11,884,671,784	111,743,279
13.	pl	8,213,686,176	107,224,899
14.	fr-fr	18,900,680,340	12,299,715
15.	lt-lt	967,974,884	9,963,470
16.	fr-ca	944,108,012	7,845,659
17.	lv-lv	1,010,410,672	6,585,033
18.	nl	16,051,153,272	5,036,266
19.	en-gb	2,622,369,340	4,993,864
20.	it-it	1,050,261,400	4,336,673

KantanMT.com

No Hardware. No Software. No Hassle MT.

Thank You!

Tonyod@KantanMT.com

KantanMT.com

No Hardware. No Software. No Hassle MT.

Thank You!

Tonyod@KantanMT.com