Translation for the rest of us:
Usable word-level translation for automated text analysis

Appendix

A. Maurits van der Veen (maurits@wm.edu)
William & Mary

Current version: 5 May 2022

Contents

1. Dictionary composition

Table 2 in the paper provides an overview, by language, of the share of translations that takes place directly and through word segmentation, respectively. Table A1 below provides more detailed information, by language, of the different means through which each word in the dictionary was translated. The values listed are percentages, and the languages are listed by language code, in the same order as in the table: Romance languages first, then Germanic languages, and finally Finnish and Polish.

The percentages listed are rounded to 2 decimal places. Values of 0.00 have been removed for greater legibility. Note, however, that these categories may still include a very small number of words translated (just rounded down to 0.00%). For a few languages, a small fraction of translations is listed as 'trusted': these are taken from the website www.101languages.net, where up to 2,000 word pairs are listed under "most common words" for each language.

Among the languages listed here, Finnish is the most difficult language to translate, given the heuristics used: as the bottom line in the table shows, nearly one quarter of all words are simply copied to the target language because no suitable translation could be found. This is due to the fact that Finnish word structure is more complex than that in most languages, in particular with

word endings that regularly exceed 4 characters. Finnish may thus be a case where an approach based on character n-grams would pay dividends (Bojanowski et al., 2016; see also Sennrich et al., 2015).

Category	pt	es	it	fr	de	nl	da	fi	pl
Trusted		1.59	1.65	1.69		0.90			
Direct	89.85	44.76	47.43	42.64	44.85	30.48	98.64	39.59	66.03
Direct, lower-casing (lc)		28.27	29.34	33.04	32.03	24.98		10.69	17.58
Accent-stripped	0.14	0.11	0.13	0.05	0.01	0.07	0.01		
Accent-stripped, lc		0.01	0.02	0.01		0.02			
Spell-checked		2.08	1.82	1.75	0.56	2.11		0.26	0.85
Spell-checked, lc		0.55	0.60	0.76	0.20	1.18		0.08	0.23
Segmented on hyphen		0.21	0.60	1.95	0.24	1.38		1.02	0.65
Segmented on hyphen, lc		0.60	0.53	1.88	6.28	2.54		1.52	0.43
Segmented on capitalization, lc		0.16	0.17	0.11	0.05	0.17			0.01
Segmented	2.27	4.54	5.02	3.20	2.79	18.12	0.09	13.12	3.19
Segmented, accent-stripped	0.07	0.48	0.07	0.16	0.10	0.03		2.10	0.07
Add/cut ending character	1.80	0.94	0.41	0.43	0.11	0.26	0.10	0.00	0.26
Add/cut ending, lc		0.05	0.05	0.07	0.02	0.11			0.03
Stemmed (up to 4 chars)	0.42	1.72	1.16	0.71	0.72	1.15		6.32	3.06
No translation found, copied	0.95	5.71	4.34	6.17	9.60	9.95	0.01	24.19	5.78

Table A1. Percentage of words translated, by type of translation

2. Sentiment analysis validation

Figure 2 in the paper displays sentiment analysis performance, by language, for the *Sourcespeeches* corpus. Table A2 compares this performance to that on the other two corpora.

	Match polarity			Correlation		
Language	***Full***	***Filtered***	***Source***	***Full***	***Filtered***	***Source***
Danish	0.84	0.83	0.83	0.81	0.79	0.77
German	0.82	0.82	0.79	0.75	0.72	0.68
Spanish	0.85	0.85	0.86	0.83	0.81	0.81
Finnish	0.81	0.81	0.84	0.73	0.70	0.76
French	0.84	0.83	0.81	0.79	0.77	0.77
Italian	0.84	0.84	0.84	0.80	0.78	0.80
Dutch	0.81	0.81	0.79	0.78	0.75	0.76
Polish	0.80	0.80	0.80	0.71	0.67	0.75
Portuguese	0.85	0.84	0.85	0.81	0.77	0.84

Average	*0.83*	*0.83*	*0.82*	*0.78*	*0.75*	*0.77*

Table A2. Sentiment analysis performance, across corpora.

3. Emotion measurement validation

Figure 3 in the paper displays emotion measurement performance, by language, for the *Sourcespeeches* corpus. In Table A3, I break down those scores by emotion, along with the overall data shown in Figure 3. Below that are the analogous scores for the *Fullspeeches* (Table A4) and *Filteredspeeches* (Table A5) corpora.

Language	**Anger**	**Anticip.**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Trust**	**Avg.**	**Cos.**
Danish	0.34	0.71	0.47	0.90	0.73	0.75	0.27	0.60	0.60	0.74
German	0.48	0.72	0.53	0.74	0.71	0.72	0.54	0.78	0.65	0.83
Spanish	0.53	0.79	0.57	0.91	0.87	0.70	0.59	0.80	0.72	0.86
Finnish	0.23	0.63	0.27	0.74	0.77	0.67	0.33	0.79	0.55	0.81
French	0.59	0.79	0.68	0.86	0.81	0.59	0.66	0.82	0.72	0.84
Italian	0.55	0.75	0.58	0.87	0.80	0.79	0.51	0.77	0.70	0.84
Dutch	0.42	0.66	0.35	0.80	0.68	0.74	0.46	0.69	0.60	0.77
Polish	0.31	0.52	0.56	0.74	0.60	0.78	0.38	0.60	0.56	0.78
Portuguese	0.45	0.77	0.56	0.91	0.80	0.79	0.54	0.79	0.70	0.86
Average	*0.43*	*0.71*	*0.51*	*0.83*	*0.75*	*0.73*	*0.48*	*0.74*	*0.65*	*0.81*

Table A3. Emotion measurement performance on *Sourcespeeches* corpus (used for Figure 3): correlation by emotion, overall, and cosine similarity across the 8-emotion vector.

Language	**Anger**	**Anticip.**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Trust**	**Avg.**	**Cos.**
Danish	0.51	0.76	0.50	0.80	0.67	0.65	0.25	0.67	0.60	0.79
German	0.44	0.71	0.57	0.75	0.67	0.70	0.55	0.72	0.64	0.83
Spanish	0.46	0.76	0.57	0.83	0.76	0.68	0.59	0.76	0.68	0.85
Finnish	0.30	0.61	0.35	0.71	0.72	0.65	0.31	0.71	0.55	0.80
French	0.56	0.76	0.70	0.81	0.71	0.59	0.62	0.76	0.69	0.86
Italian	0.53	0.74	0.52	0.81	0.73	0.76	0.56	0.74	0.67	0.84
Dutch	0.40	0.64	0.39	0.76	0.69	0.74	0.43	0.69	0.59	0.79
Polish	0.31	0.60	0.51	0.70	0.63	0.67	0.40	0.71	0.57	0.79
Portuguese	0.52	0.77	0.62	0.84	0.72	0.79	0.61	0.78	0.71	0.87
Average	*0.45*	*0.71*	*0.52*	*0.78*	*0.70*	*0.69*	*0.48*	*0.73*	*0.63*	*0.83*

Table A4. Emotion measurement performance on *Fullspeeches* corpus:
correlation by emotion, overall, and cosine similarity across the 8-emotion vector.

Language	Anger	Anticip.	Disgust	Fear	Joy	Sadness	Surprise	Trust	Avg.	Cos.
Danish	0.51	0.76	0.50	0.79	0.67	0.65	0.25	0.67	0.60	0.80
German	0.44	0.71	0.56	0.74	0.67	0.68	0.55	0.71	0.63	0.84
Spanish	0.47	0.76	0.59	0.82	0.76	0.67	0.59	0.75	0.68	0.86
Finnish	0.29	0.61	0.34	0.70	0.72	0.63	0.30	0.71	0.54	0.81
French	0.56	0.76	0.69	0.80	0.71	0.58	0.62	0.74	0.68	0.87
Italian	0.52	0.74	0.50	0.80	0.73	0.75	0.55	0.73	0.66	0.85
Dutch	0.39	0.64	0.37	0.75	0.69	0.73	0.43	0.68	0.58	0.81
Polish	0.31	0.60	0.51	0.68	0.63	0.67	0.39	0.70	0.56	0.81
Portuguese	0.51	0.77	0.61	0.83	0.72	0.79	0.60	0.77	0.70	0.87
Average	*0.44*	*0.71*	*0.52*	*0.77*	*0.70*	*0.68*	*0.48*	*0.72*	*0.63*	*0.84*

Table A5. Emotion measurement performance on *Filteredspeeches* corpus: correlation by emotion, overall, and cosine similarity across the 8-emotion vector.

4. Topic modeling validation

Figures 4 and 5 in the text show the similarity in topic weights between professional human translations and the machine-translated texts. Table A6 lists the data used to generate these figures, along with the analogous data for the *Filteredspeeches* corpus.

	Cosine similarity			Correlation		
Language	***Full***	***Filtered***	***Source***	***Full***	***Filtered***	***Source***
Danish	0.87	0.83	0.77	0.82	0.76	0.70
German	0.86	0.82	0.75	0.82	0.76	0.70
Spanish	0.93	0.92	0.85	0.91	0.90	0.83
Finnish	0.83	0.79	0.78	0.79	0.71	0.68
French	0.91	0.90	0.83	0.89	0.88	0.79
Italian	0.90	0.90	0.84	0.88	0.87	0.80
Dutch	0.89	0.84	0.80	0.87	0.80	0.76
Polish	0.90	0.84	0.81	0.87	0.80	0.75
Portuguese	0.93	0.92	0.86	0.90	0.89	0.79
Average	*0.89*	*0.86*	*0.81*	*0.86*	*0.82*	*0.76*

Table A6. Sentiment analysis performance, across corpora.

In addition, Tables A7 and A8 show, for the *Fullspeeches* and *Filteredspeeches* corpora, respectively, the full lists of topics generated, both substantive and non-substantive. Note that the substantive topics are basically the same for the two corpora.

Topic	Top 6 words
Economics & welfare	economic, social, crisis, financial, development, growth
	ones, chairman, exceptionally, rapporteur, mandates, entering
	dear, amending, considered, determination, subcommittee, donate
	needs, ussr, vice, wants, enter, lies
	rhi, promotable, rapporteur, speaker, consultation, favor
Rights	rights, human, country, international, freedom, democracy
	proper, order, everybody, eu, parliamentary, lord
	ramble, ends, saw, monsieur, discrepancy, united
EU institutions	commission, council, parliamentx, day, order, proposal
	states, sole, honorable, president, want, relationship
Legislative action	report, voted, favor, written, vote, eu
Voting	voting, article, debates, closed, 142, 149
Women & children	women, men, equality, violence, children, social
EU markets	products, consumers, market, protection, health, agricultural
	lord, united, vice, guess, kind, wanted
Energy	energy, nuclear, climate, efficiency, emissions, renewables

Table A7. Topics in *Fullspeeches* corpus (non-substantive, or 'junk' topics grayed out). 8 out of 16 topics are substantively meaningful.

Topic	Top 6 words
Member states	members, states, united, member, union, national
	ones, rapporteur, exceptionally, mandates, entering, chairman
	dear, amending, considered, determination, subcommittee, donate
	needs, ussr, wants, enter, lies, nations
	rhi, rapporteur, promotable, speaker, consultation, hope
EU markets	products, consumers, market, protection, agricultural, production
	proper, order, everybody, eu, parliamentary, da
EU institutions	commission, council, parliamentx, proposal, day, order
	ramble, ends, saw, discrepancy, monsieur, political
Women & children	women, men, equality, violence, children, social
	sole, honorable, relationship, want, doing, breaks
Rights	rights, human, country, international, democracy, freedom
Economics & welfare	economic, social, crisis, financial, development, growth
	sole, lord, mi, señora, president, ocurred
Energy	energy, nuclear, climate, efficiency, renewables, gas
	vice, lord, guess, kind, wanted, stands
Legislative action	voted, report, favor, written, resolution, vote
	voting, march, thursday, place, 00, 12

Table A8. Topics in *Filteredspeeches* corpus (non-substantive, or 'junk' topics grayed out). 8 out of 18 topics are substantively meaningful.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching Word Vectors with Subword Information*. Arxiv. https://doi.org/10.48550/ARXIV.1607.04606

Sennrich, R., Haddow, B., & Birch, A. (2015). *Neural Machine Translation of Rare Words with Subword Units*. Arxiv. https://doi.org/10.48550/arXiv.1508.07909