Annex A: Systematization of several available corpus.

Name	Authors	Number of languages	Number of documents per language	Licensing	Original objective of the corpus	Main source of data
MARC	Keung et al. 2020	6	210 thousand	Open access. Custom license - amazon.com terms of use	Classification according to the number of stars.	Amazon Reviews Customer
MuST-C	(Cattoni et al. 2020)	14	270 thousand	CC BY-NC-ND	Machine translation tasks.	TED talks
Multilingual Corpus of Online Educational Content	(Sosoni et al. 2018)	11	87 thousand	Custom: h2020 Open Research Data Pilot	Machine translation tasks.	Iversity.org Videolectures.net Coursera QED
SNLI	(Bowman et al. 2015)	1	570 thousand	CC BY-ND	Development and evaluation of models for understanding sentences.	Flickr30k corpus
XNLI	(Conneau et al. 2018)	15	10 thousand	Several, depending on the data group: CC BY-ND Custom: OANC	Development and evaluation of multilingual models for understanding sentences.	MultiNLI
MultiNLI	Williams, Nangia & Bowman (2018)	1	433 thousand	Several, depending on the data group: CC BY-ND Custom: OANC.	Development and evaluation of models for understanding sentences.	Open American National Corpus (OANC)
Yelp Open Dataset	(Yelp 2019)	Various. Not specified.	8 million in total	Apache-2.0	Academic and research purposes	Platform user comments
Amazon Customer Reviews	(Amazon 2020)	Various. Not specified.	130 million in total	Open access. Custom license - amazon.com terms of use	Academic and research purposes	Platform user comments
RCV2	(Reuters 2019)	13	487 thousand	Access: private. License: under Reuters News Terms of Use	Research development in NLP	Reuters News Articles
MEANTIME	Minard et al. (2016)	4	120	CC-BY	Automatic text generation	Portal Wikinews
Europarl	Koehn (2005)	11	Varies from 200 thousand to 2 million	CC0	Machine translation tasks.	Acts of the European parliaments
MLSUM	Scialom et al. (2020)	5	200 thousand	Unknown	Automatic text summarization tasks.	News web portals
MLQA	Lewis et al. (2019)	7	12 thousand in English y 5 thousand for each other language	CC BY-NC	Support for the development of question and answer systems.	Wikipedia
XTREME	Hu et. al (2020)	40	15 GB in total	Several, depending on the data group: CC-BY CC BY-NC CC BY-ND	Benchmarks in 9 tasks	Multiple large-scale corpus.

XGLUE	Liang et al. (2020)	89	2 TB in total	Several, depending on the data group: CC-BY CC BY-NC CC BY-ND	Benchmarks in 11 tasks	Multiple large-scale corpus.