CONTRAST-IT SELECT. A highly comparable corpus

In light of the central role played by language comparability in contrastive and comparative linguistics, the **CONTRAST-IT SELECT** corpus has been devised based on the same data collection used in the creation of the CONTRAST-IT general corpus.

The CONTRAST-IT-SELECT corpus is made of highly comparable language subcorpora: each of them includes the three same thematic sections (politics, economy and sports) and each thematic section has the same size (50’000 words). The corpus design is presented in Table 2.

Table 2. CONTRAST-IT SELECT. Corpus design (thematic sections and size)

	IT	FR	SP	E	G
Politics	50'000	50'000	50'000	50'000	50'000
Economy	50'000	50'000	50'000	50'000	50'000
Sports	50'000	50'000	50'000	50'000	50'000

Below is a detailed description of the corpus design of CONTRAST-IT SELECT. Information is given on the thematic sections of the news, the number of words collected for each thematic section and the dates of publication of the articles.

CONTRAST-IT SELECT_IT (Italy): 150’000 words

Politics (Politica):

corriere.it	20'000	11.2011 – 12.2011
lastampa.it	20'000	11.2011
repubblica.it	10'000	05.2011 – 11.2011

Economy (Economia, finanza):

corriere.it	30'000	09.2011 – 12.2011
repubblica.it	20'000	09.2011 – 10.2011

Sports (Sport):

corriere.it	20'000	01.2011 – 12.2011
lastampa.it	10'000	01.2011 – 12.2011
repubblica.it	20'000	03.2011 – 12.2011

CONTRAST-IT SELECT_FR (France): 150’000 words

Politics (Politique):

lefigaro.fr	25'000	07.2011 – 12.2011
lemonde.fr	25'000	01.2011 – 11.2011

Economy (Économie):

lefigaro.fr	20'000	12.2011
lemonde.fr	30'000	10.2011 – 11.2011

Sports (Sport):

lefigaro.fr	30'000	10.2011 – 12.2011
lemonde.fr	20'000	08.2011 – 11.2011

CONTRAST-IT SELECT_SP (Spain): 150’000 words

Politics (Política):

elpais.com	50'000	09.2011 – 03.2012

Economy (Economía):

elpais.com	35'000	10.2011 – 03.2012
elmundo.es	15'000	02.2011 – 08.2011 and 03.2012

Sports (Deportes):

elpais.com	25'000	02.2012 – 03.2012
elmundo.es	25'000	01.2011 – 03.2012

CONTRAST-IT SELECT_E (England): 150’000 words

Politics (News_Politics, Politics):

theguardian.com	10'000	10.2011 – 11.2011
independent.co.uk	40'000	02.2011 – 12.2011

Economy (News_Business, Business):

theguardian.com	25'000	07.2011 – 11.2011
independent.co.uk	25'000	01.2011 – 03.2011 and 01.2015

Sports (Sport, Sports):

theguardian.com	10'000	04.2011 – 12.2011
independent.co.uk	40'000	12.2014 – 02.2015

CONTRAST-IT SELECT_G (Germany): 150’000 words

Politics (Politik):

faz.net	25'000	11.2011
sueddeutsche.de	25'000	02.2011 – 12.2011

Economy (Wirtschaft, Wirtschaft & Geld):

faz.net	30'000	09.2011 – 12.2011
sueddeutsche.de	20'000	12.2011

Sports (Sport):

faz.net	25'000	11.2011– 12.2011
sueddeutsche.de	25'000	02.2011 – 12.2011