

COMPARE-IT. Detailed corpus design

Here is a detailed description of the corpus COMPARE-IT. Information is given on the thematic sections of the articles collected, the number of words collected in each thematic section and the dates of publication of the articles.

COMPARE-IT_IT (Italy): 300'000 words

repubblica.it (165'000)

- Cronaca, 71'813 words, 08.2011 – 10.2011
- Sport, 20'323 words, 03.2011 – 12.2011
- Tecnologia, 20'031 words, 09.2011 – 10.2011
- Economia e finanza, 18'522 words, 09.2011 – 10.2011
- Esteri, 11'970 words, 09.2011 – 11.2011
- Politica, 10'445 words, 05.2011 – 11.2011
- Scienze, 7'701 words, 09.2011 – 10.2011
- Pubblico, 5'007 words, 11.2011

corriere.it (85'000)

- Economia, 31'622 words, 09.2011 – 12.2011
- Sport, 19'641 words, 01.2011 – 12.2011
- Politica, 19'454 words, 11.2011 – 12.2011
- Cronaca, 15'122 words, 10.2011

lastampa.it (50'000)

- Cronache, 20'719 words, 11.2011
- Politica, 20'326 words, 11.2011
- Sport, 10'361 words, 01.2011 – 12.2011

COMPARE-IT_CH (Switzerland): 150'000 words

cdt.ch (130'000)

- Politica, 50'682 words, 01.2013 – 12.2013
- Economia, 43'479 words, 01.2013 – 12.2013
- Sport, 39'654 words, 01.2013 – 12.2013

gdp.ch (20'000)

- Sport, 11'186 words, 08.2013 – 09.2013
- Economia, 6'884 words, 03.2013 – 09.2013

COMPARE-IT_CA (Canada): 105'000 words

corriere.ca (105'000)

- Canada, 80'906 words, 06.2012 – 05.2013
- Comunità, 15'497 words, 09.2012 – 05.2013
- Cultura, 3'294 words, 07.2012 – 11.2012
- Italia, 2'676 words, 02.2013 – 03.2013
- Primo piano, 1'218 words, 11.2012
- Spettacoli, 1'080 words, 10.2012
- Sport, 630 words, 04.2013
- Focus, 464 words, 03.2013