CLARIN RESOURCE FAMILIES

FULL PAPER:
Darja Fišer [1,2] Jakob Lenardič [1] Tomaž Erjavec [2]
[1] Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia
[2] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
{darja.fiser, jakob.lenardic}@ff.uni-lj.si, tomaz.erjavec@ijs.si

For more information visit: www.clarin.eu/resource-families

NEWSPAPER CORPORA
27 SURVEYED

4 MULTILINGUAL
23 MONOLINGUAL: 8 LANGUAGES

1 Arabic, 2 Czech, 1 Finnish, 2 French, 4 German, 1 Norwegian, 1 Polish, 11 Swedish

AVAILABILITY
5 through a concordancer
10 for download
11 both

ANNOTATION
12 PoS-tagged
4 lemmatised

SIZE
11 small (<10 million tokens)
3 medium (10–100 million tokens)
6 large (>100 million tokens)

LICENCE
12 CC-BY
4 ELRA END USER/VAR
2 CLARIN PUB
1 CLARIN ACA

PARLIAMENTARY CORPORA
16 SURVEYED

1 MULTILINGUAL
15 MONOLINGUAL: 13 LANGUAGES

1 Czech, 1 Danish, 2 English, 1 French, 1 Estonian, 1 Finnish, 2 German, 1 Greek, 1 Lithuanian, 2 Norwegian, 1 Portuguese, 1 Slovenian, 1 Swedish

AVAILABILITY
3 through a concordancer
9 for download
4 both

ANNOTATION
7 PoS-tagged
7 lemmatised

SIZE
7 small (<10 million tokens)
6 medium (10–100 million tokens)
2 large (>100 million tokens)

LICENCE
8 CC-BY
1 CLARIN ACA
1 CLARIN PUB

PARALLEL CORPORA
81 SURVEYED

47 BILINGUAL:
mostly European but also Hindi, Tamil and Vietnamese

34 MULTILINGUAL:
5 contain over 50 languages

AVAILABILITY
5 through a concordancer
57 for download
4 both

ANNOTATION
38 sentence-aligned
4 word-aligned
3 paragraph-aligned

SIZE
29 small (<10 million tokens)
13 medium (10–100 million tokens)
8 large (>100 million tokens)

LICENCE
30 CC-BY
10 ELRA END USER/VAR
10 open for reuse with restrictions

CMC CORPORA
12 SURVEYED

8 LANGUAGES:

1 Czech, 1 Dutch, 1 Estonian, 1 Finnish, 1 French, 1 German, 1 Lithuanian, 5 Slovenian

AVAILABILITY
1 through a concordancer
4 for download
7 both

ANNOTATION
8 PoS-tagged
6 lemmatised
5 word-normalised

SIZE
4 small (<10 million tokens)
6 medium (10–100 million tokens)
2 large (>100 million tokens)

LICENCE
8 CC-BY
2 CLARIN ACA

L2 LEARNER CORPORA
34 SURVEYED

10 MULTILINGUAL
24 MONOLINGUAL: 9 LANGUAGES

1 Arabic, 1 Czech, 10 English, 4 Finnish, 1 French, 2 German, 1 Hungarian, 1 Norwegian, 3 Swedish

AVAILABILITY
5 through a concordancer
15 for download
3 both

ANNOTATION
5 PoS-tagged
1 lemmatised

SIZE
8 small (<10 million tokens)
5 medium (10–100 million tokens)
0 large (>100 million tokens)

LICENCE
12 CLARIN RES
10 CC-BY
2 ELRA END USER/VAR