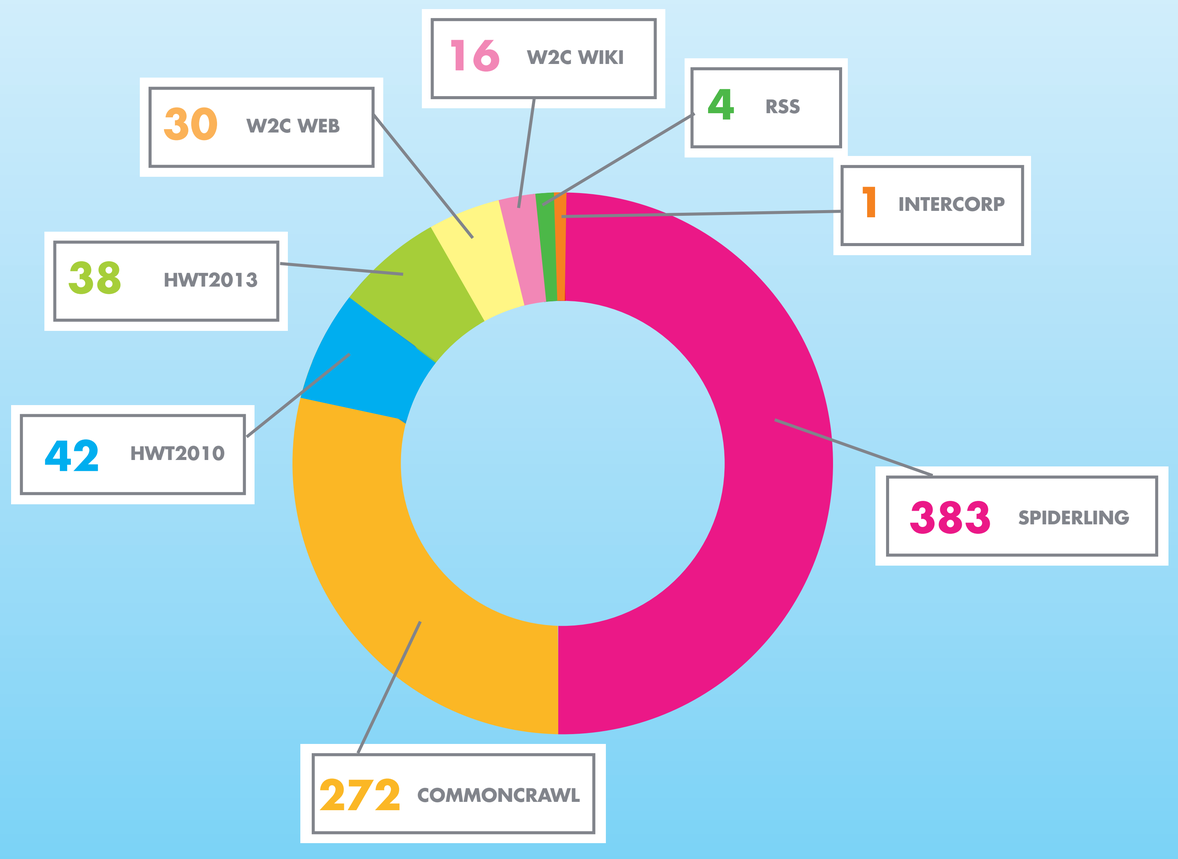
HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation

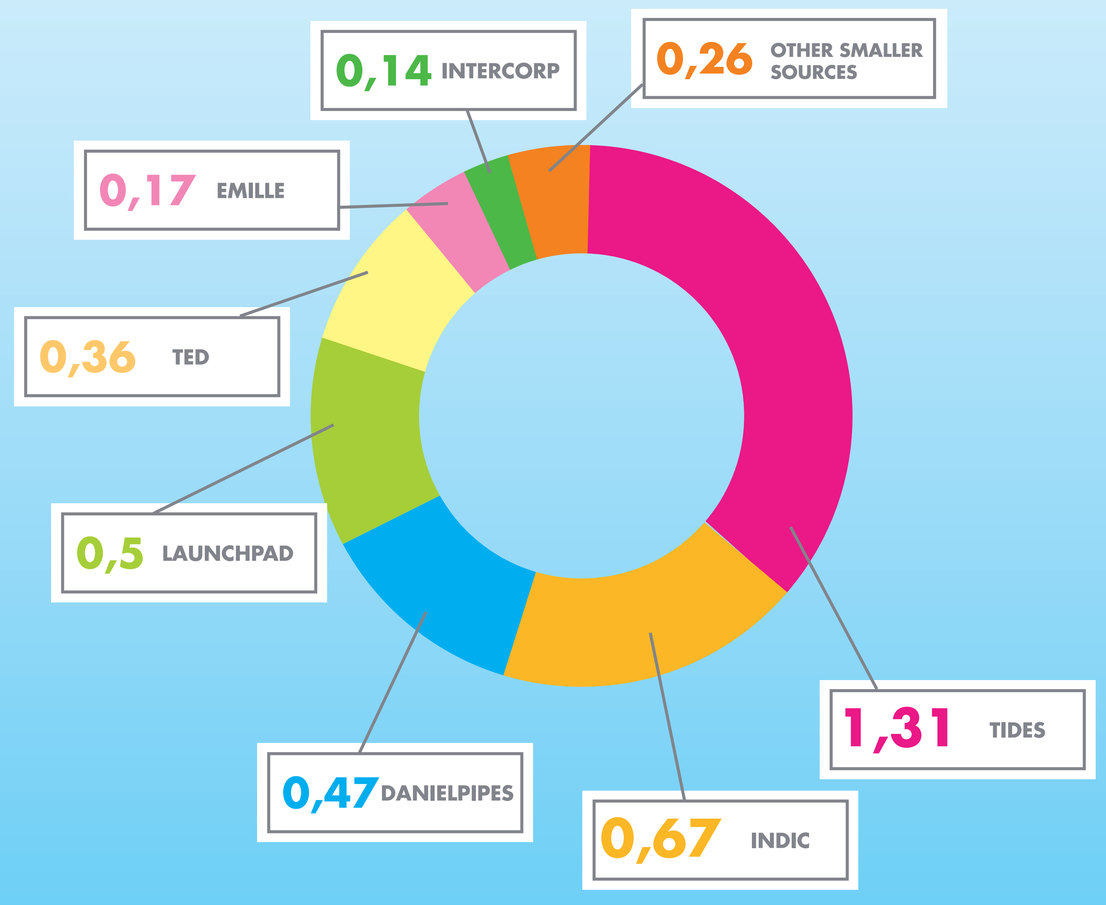
Ondřej Bojar[1], Vojtěch Diatka[2], Pavel Rychlý[3], Pavel Straňák[1], Vít Suchomel[3], Aleš Tamchyna[1], Daniel Zeman[1]

1 - Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
2 - Charles University in Prague, Faculty of Arts, Department of Linguistics
3 - Natural Language Processing Centre, Faculty of Informatics, Masaryk University

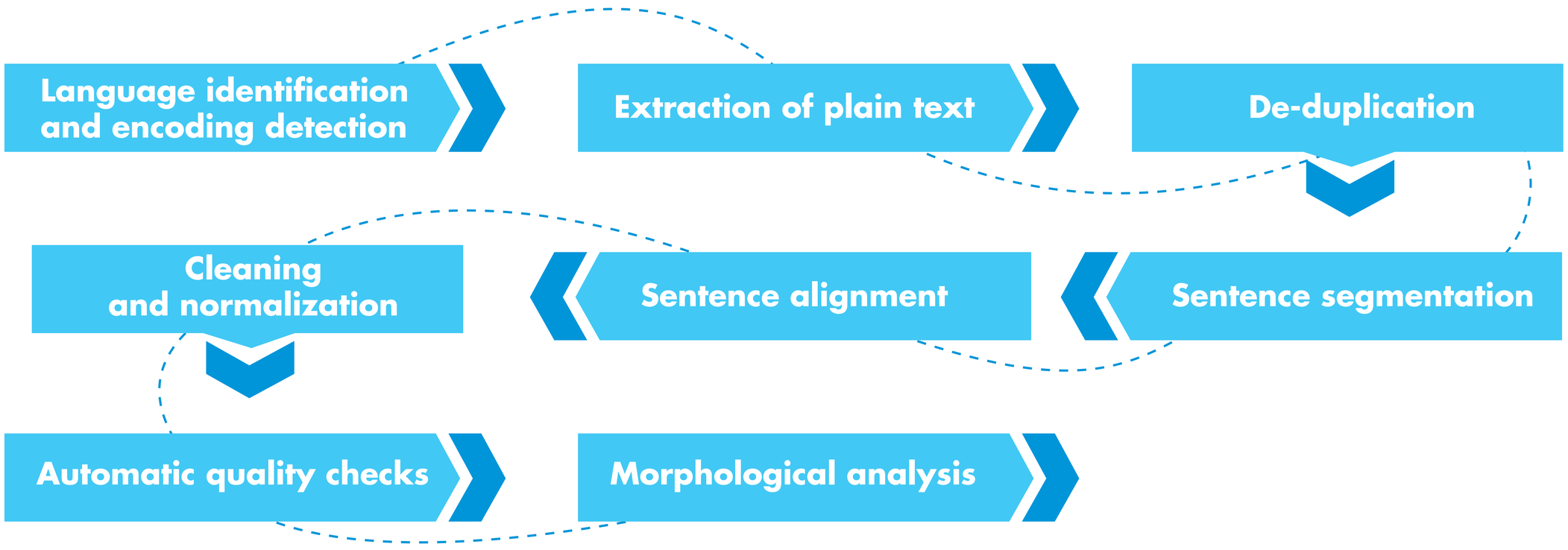
HindEnCorp 0.5 sections and statistics



HindMonoCorp 0.5 sections and statistics



PROCESSING PIPELINE



CLEANING AND NORMALIZATION

- various typesetting conventions and typesetting errors
- resolving some prominent character-level inconsistencies
  - removing various non-printable chars
  - normalizing Unicode
  - correcting typesetting of Devanagari nukta
  - danda vs full stop
  - arabic vs Devanagari numbers

उत्पन्न (970) seen as उत्पन्न 63.8% उत्पऩ्न 35.2% उत्पन्न 0.6% उत्पऩ्न 0.3% उत्पन्न 0.1%

व्यक्तियों (503) seen as व्यक्तियों 71.0% व़्यक़्तियों 15.3% व़्यक्तियों 12.3% व़्यक़्तियों 1.0% व़्यक्तियों 0.2% व्यक्तियों 0.2%

बल्लेबाज (489) seen as बल्लेबाज 93.7% बल़्लेबाज 5.1% बल्लेबाज़ 1.0% बल़्लेबाज़ 0.2%

अस्तित्व (407) seen as अस्तित्व 82.1% अस़्तित़्व 13.5% अस्तित़्व 4.4%

अफगानिस्तान (336) seen as अफगानिस्तान 87.2% अफ़ग़ानिस्तान 7.1% अफ़गानिस्तान 3.0% अफगानिस़्तान 2.7%

मस्तिष्क(321) seen as मस्तिष्क 70.4% मस़्तिष़्क 26.5% मस्तिष़्क 1.2% मस़्तिष्क 1.2% मस़्तिष़्क 0.3% मस्तिष्क़ 0.3%

STATISTICS ON CLEANING AND TYPOGRAPHICAL

	CommonCra	*SpiderLing*	*HWT*	*W2C Web*	*W2C Wiki*	*RSS*	*Launchpad*	*TIDES*	*WikiNE*	*TED*	*Intercorp*	*Indic*	*Emille*	*DanielPipes*	*Dictionaries*	*ACL2005*	*Agrocorpus*
nukta checked	28.1	37.7	37.4	39.6	28.0	35.7	18.1	150.9	14.5	18.6	67.1	26.8	39.6	25.8	21.5	52.6	68.8
after a bad letter dropped	0.2	0.4	0.5	0.3	0.2	0.1	0.1	96.9	0.1	0.0	0.0	0.5	0.0	0.6	0.0	0.0	0.3
after a good letter preserved	27.9	37.3	36.9	39.3	27.8	35.5	18.0	54.0	14.4	18.5	67.1	26.3	39.6	25.2	21.5	52.6	68.5
removed sequences of nuktas	0.0	0.0	0.0	0.0	0.0	0.0	7.8	0.0	0.1	0.1	0.0	0.0	103.1	0.1	9.2	0.0	0.0
danda seen	35.1	66.9	70.8	60.6	57.0	79.6	1.8	0.0	0.0	16.7	84.1	59.4	55.4	100.2	28.8	60.9	65.6
full stop seen	51.5	43.9	31.3	59.7	42.3	19.8	31.2	104.6	0.6	19.6	11.9	22.8	8.3	70.0	20.8	6.9	65.9
double danda seen	0.2	0.1	0.0	0.1	1.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
hindi digit seen	7.4	5.7	1.4	5.7	57.5	0.0	1.5	0.0	9.0	6.1	0.2	59.0	12.4	0.0	6.1	9.9	247.2
euroarabic digit seen	70.4	49.2	53.2	33.9	77.1	73.5	11.2	75.8	9.6	6.5	0.4	49.9	70.6	149.8	11.4	35.4	50.1
nbsp changed to space	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0
removed zero width joiner	0.0	0.0	0.0	0.0	0.0	0.0	3.4	0.0	0.1	0.7	0.0	2.2	0.0	0.0	0.0	0.0	0.0
# lines	20.9M	19.4M	14.4M	2.2M	812.8k	259.5k	66.7k	50.0k	46.0k	39.9k	38.4k	37.7k	10.7k	10.6k	6.5k	3.4k	0.6k

This work was supported by the grant FP7-ICT-2011-7-288487 (MosesCore) and the LINDAT/CLARIN project. Presented at LREC 2014