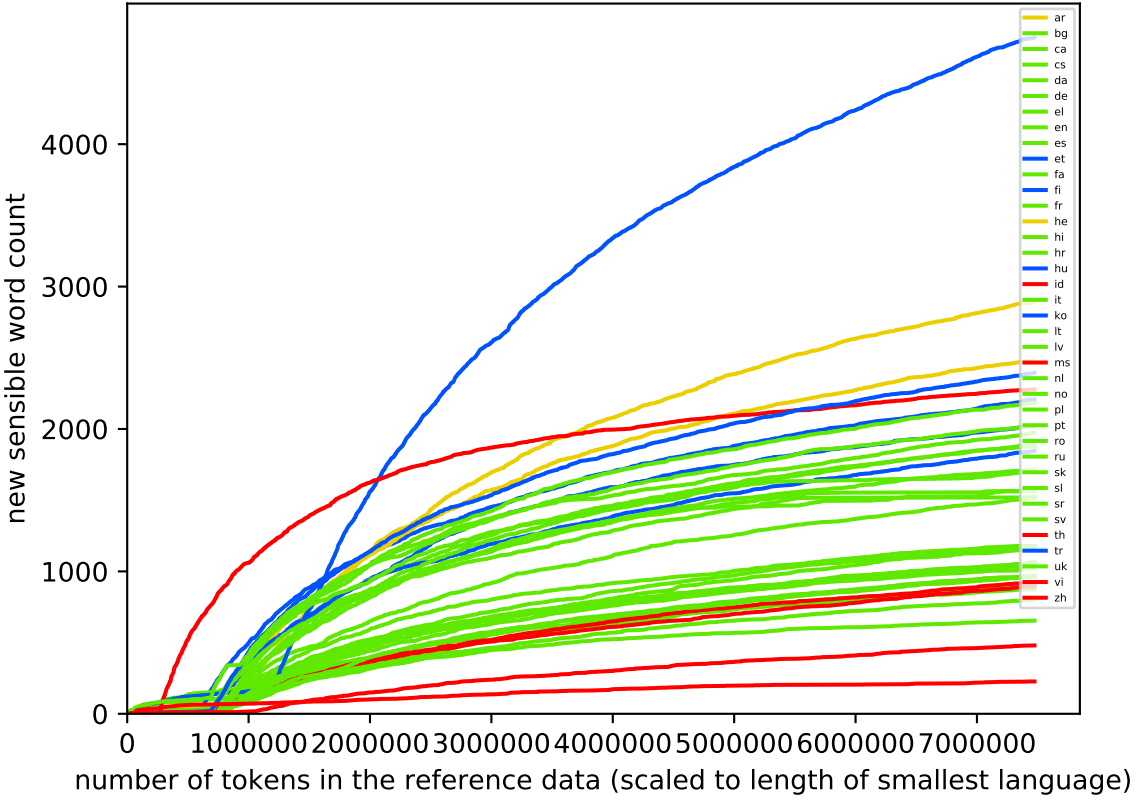
new sensible word count
0
1000
2000
3000
4000
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
number of tokens in the reference data (scaled to length of smallest language)
ar
bg
ca
cs
da
de
el
en
es
et
fa
fi
fr
he
hi
hr
hu
id
it
ko
lt
lv
ms
nl
no
pl
pt
ro
ru
sk
sl
sr
sv
th
tr
uk
vi
zh