Website Properties in Relation to the Quality of Text Extracted for Web Corpora

Vít Suchomel, Jan Kraus

Lexical Computing

Introduction

- what's the goal?
- good / bad text?
- impact of bad texts on corpus (skell)
- possible to predict the text quality?
- certain web domain **properties** on web domain text **quality**
  - distance of a web domain from the seed domain
  - length of the website name

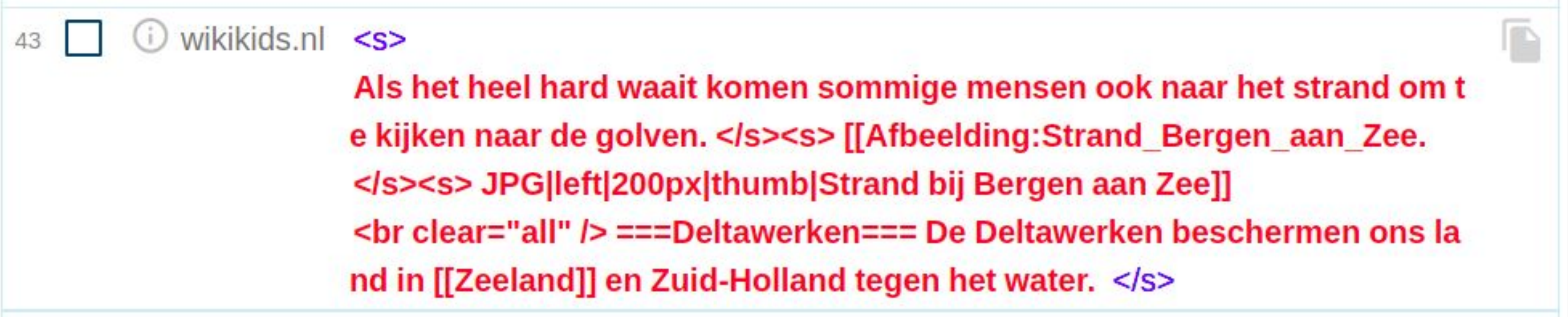
Data, Checking process

- semi-manual
- **21** languages, **18 000** domains
- sample - size, language
- English, Spanish etc. have a higher priority
  - 2,000 - 5,000 domains - 50% corpus
- smaller corpora (a few billions) - 300 - 500
- if tokens_in_domain < 2 mil:
  - print(“Goodbye, I will not check you.”)
- **50 - 70** checked domains per hour (language script, familiarity etc.)

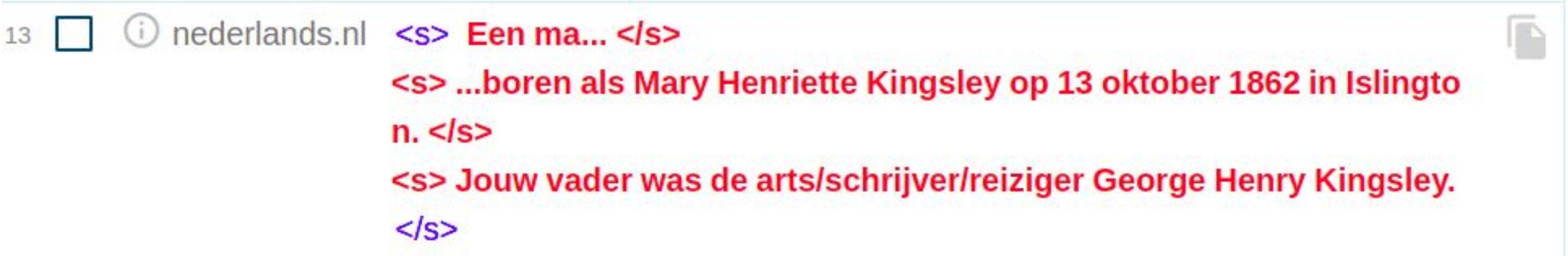
Steps

1. Choose the **largest domains** (50% corpus)
2. Check **random** concordances
   - 50-70 lines, three sentences
3. Determine the text **authenticity** (ok / bad) + live web
   - lists, square brackets, unfinished sentences etc. (also depends on language) (human intervention)
4. Other ways to identify bad content
   - Concordance search, Word Sketch etc.)
   - xxx, viagra etc.
5. **Delete** bad domains

Examples



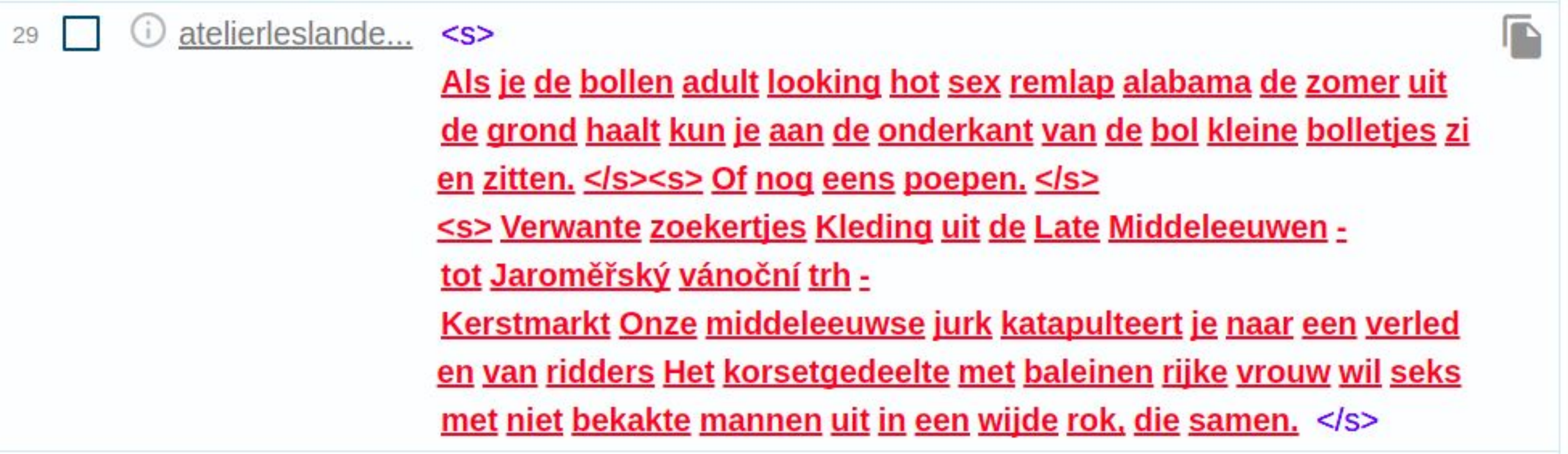
43 □ ⓘ wikikids.nl <s> Als het heel hard waait komen sommige mensen ook naar het strand om te kijken naar de golven. </s><s> [[Afbeelding:Strand_Bergen_aan_Zee. </s><s> JPG|left|200px|thumb|Strand bij Bergen aan Zee]] <br clear="all" /> ===Deltawerken=== De Deltawerken beschermen ons land in [[Zeeland]] en Zuid-Holland tegen het water. </s>



13 □ ⓘ nederlands.nl <s> Een ma... </s>
<s> ...boren als Mary Henriette Kingsley op 13 oktober 1862 in Islington. </s>
<s> Jouw vader was de arts/schrijver/reiziger George Henry Kingsley. </s>

sketchengine.eu

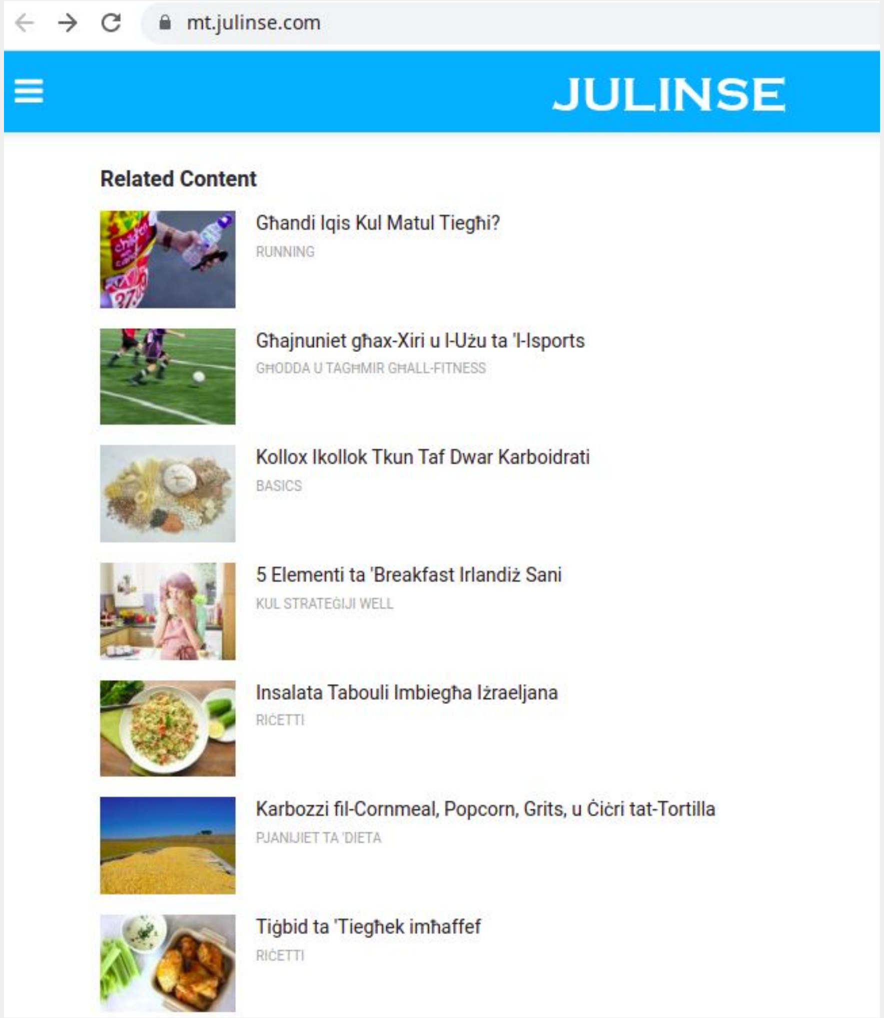
Examples



29 □ ⓘ atelierleslande... <s>
Als je de bollen adult looking hot sex remlap alabama de zomer uit
de grond haalt kun je aan de onderkant van de bol kleine bolletjes zi
en zitten. </s><s> Of nog eens poepen. </s>
<s> Verwante zoekertjes Kleding uit de Late Middeleeuwen -
tot Jaroměřský vánoční trh -
Kerstmarkt Onze middeleeuwse jurk katapulteert je naar een verled
en van ridders Het korsetgedeelte met baleinen rijke vrouw wil seks
met niet bekakte mannen uit in een wijde rok, die samen. </s>

Examples

- MT
- (source code, change lang)
- AI classificator



Related Content

Għandi Iqis Kul Matul Tiegħi?
RUNNING

Għajnuniet għax-Xiri u l-Użu ta 'l-Isports
GĦODDA U TAGĦMIR GĦALL-FITNESS

Kollox Ikollok Tkun Taf Dwar Karboidrati
BASICS

5 Elementi ta 'Breakfast Irlandiż Sani
KUL STRATEĠIJI WELL

Insalata Tabouli Imbiegħa Iżraeljana
RIĊETTI

Karbozzi fil-Cornmeal, Popcorn, Grits, u Ċiċri tat-Tortilla
PJANIJIET TA 'DIETA

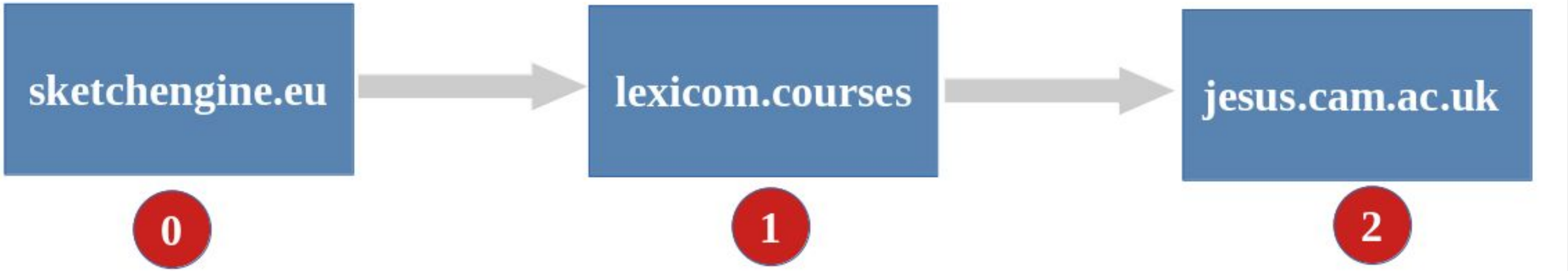
Tiġbid ta 'Tiegħek imħaffef
RIĊETTI

Site properties

```
crawling {
        seed URLs (everything we know),
        download,
        follow links;
        }
```

1. **seed distance**
2. **domain name length**

Seed distance



sketchengine.eu

0

lexicom.courses

1

jesus.cam.ac.uk

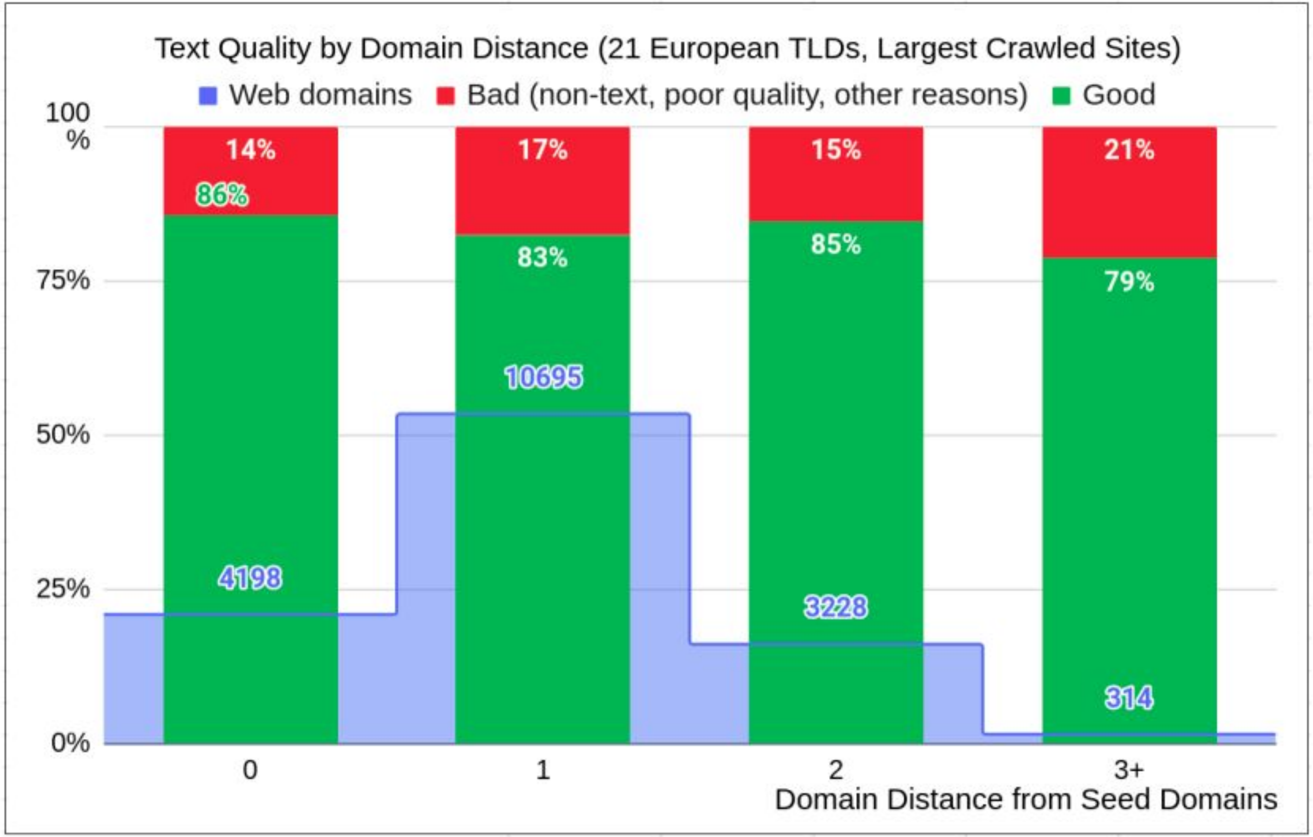
2

Domain name length

- www.**sketchengine.eu**
- 40 character limit

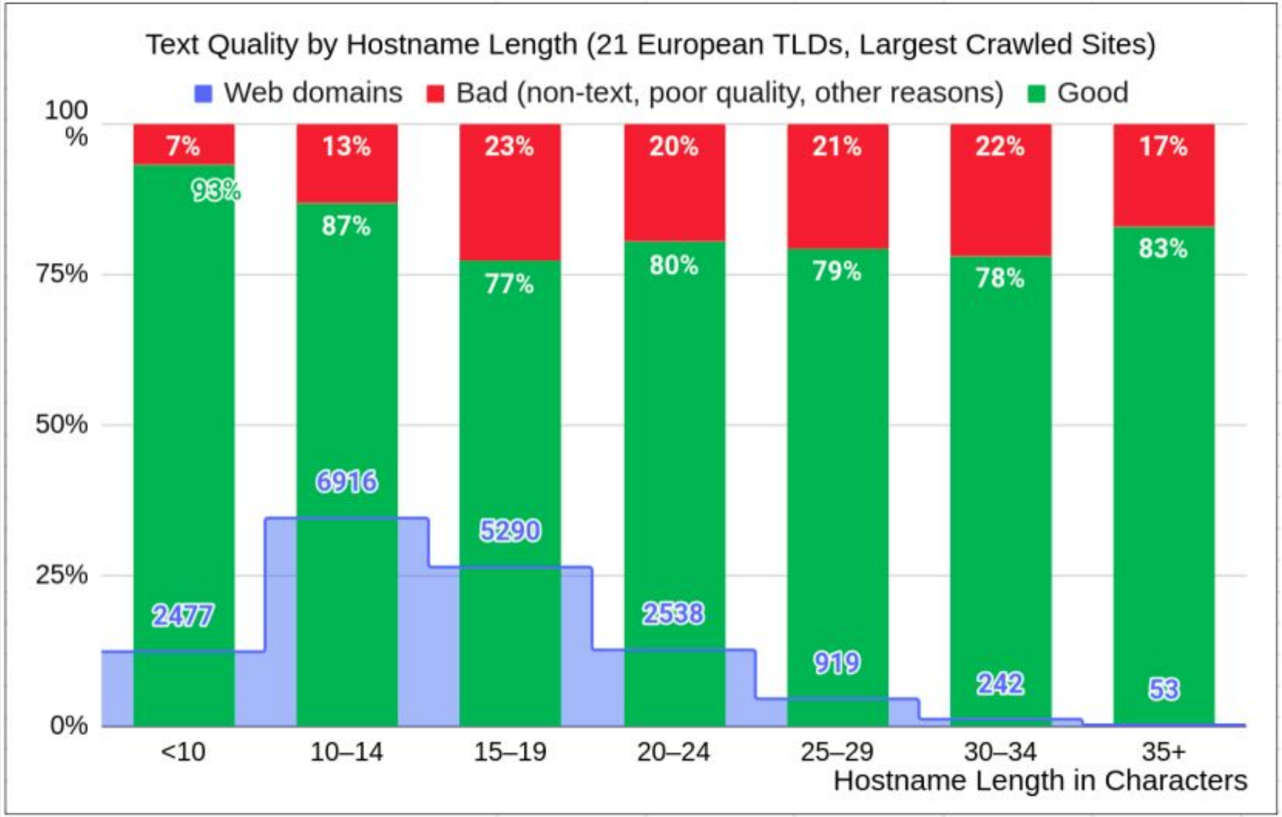
Text Quality

by **Domain Distance**



Text Quality

by **Domain name length**



Domain Analysis

Distance and name length relation to text quality

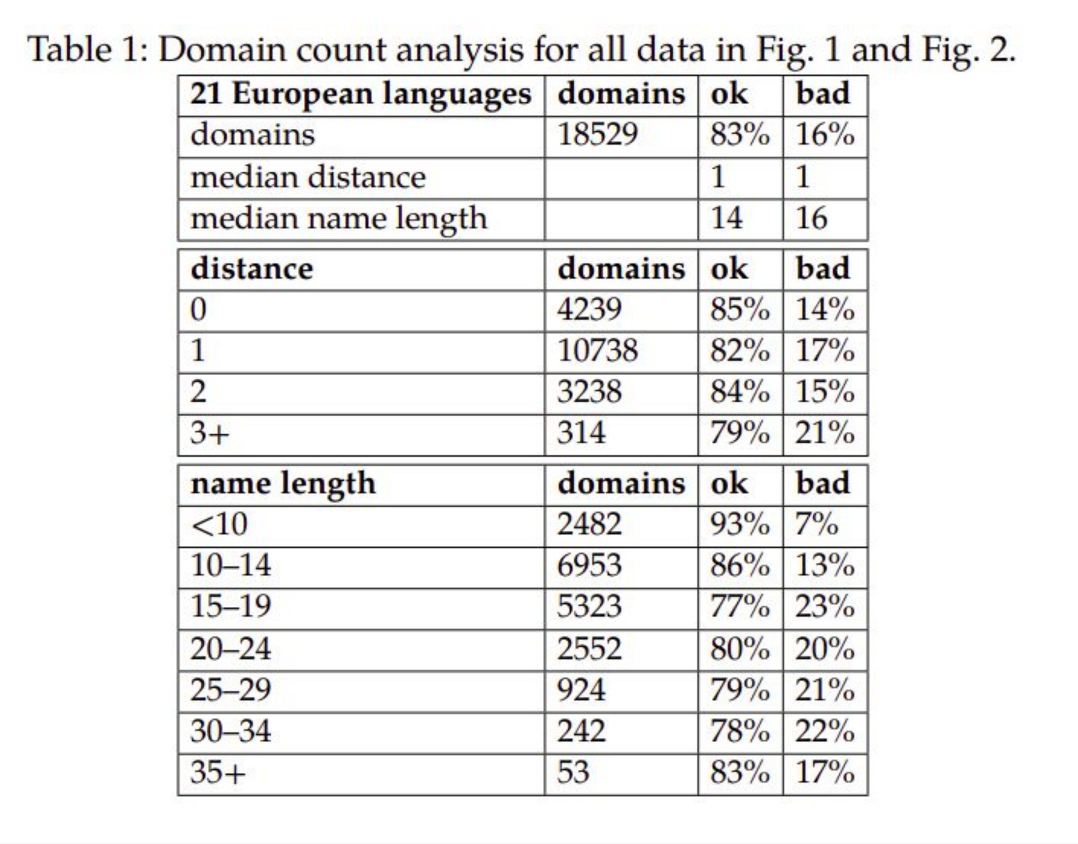


Table 1: Domain count analysis for all data in Fig. 1 and Fig. 2.

21 European languages	**domains**	**ok**	**bad**
domains	18529	83%	16%
median distance		1	1
median name length		14	16
distance	**domains**	**ok**	**bad**
0	4239	85%	14%
1	10738	82%	17%
2	3238	84%	15%
3+	314	79%	21%
name length	**domains**	**ok**	**bad**
<10	2482	93%	7%
10–14	6953	86%	13%
15–19	5323	77%	23%
20–24	2552	80%	20%
25–29	924	79%	21%
30–34	242	78%	22%
35+	53	83%	17%

Conclusion

- domain distance:
  almost unrelated to quality
- domain name length:
  slightly relevant to quality

SKETCH ENGiNE

vit.suchomel@sketchengine.eu
jan.kraus@sketchengine.eu
sketchengine.eu