

Report from Crowdfest

Crowdsourcing corpus cleaning for language learning

(Spin-off idea from WG1 hands-on meeting)

Tanara Zingano Kuhn, CELGA-ILTEC, University of Coimbra
Peter Dekker, Dutch Language Institute
March, 14th 2019, Lisbon



enetCollect

COST Action CA16105

<http://enetcollect.eurac.edu/>

enetcollect@gmail.si



COST is supported by the EU
Framework Programme
Horizon 2020

The hardworking team



Overview

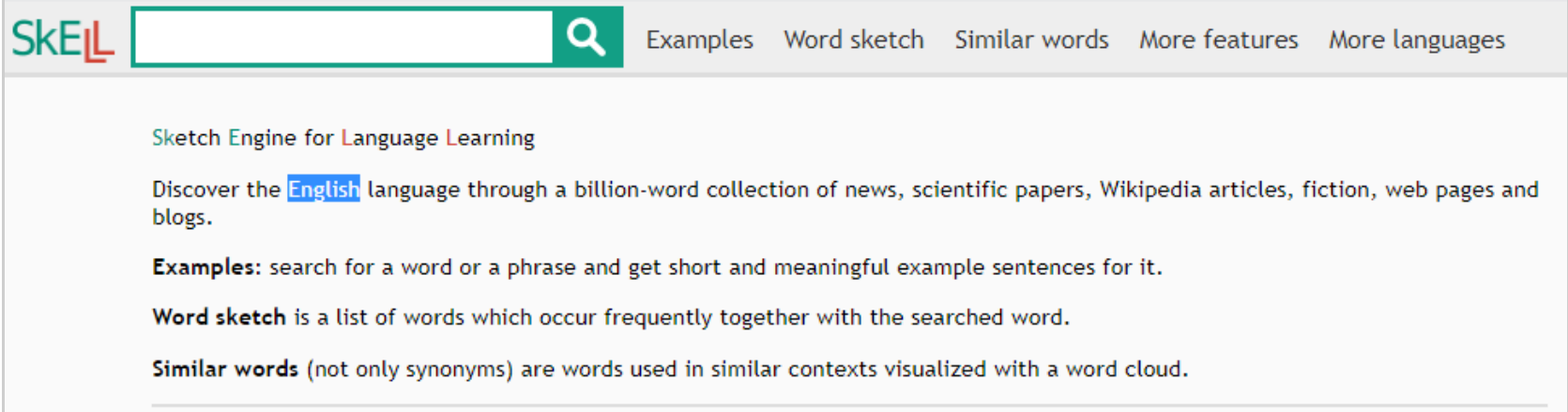
- A good (tentative) idea
- A hardworking group
- What's next?



The ptSkELL project

This project sets up to develop the Portuguese version of SkELL (Sketch Engine for Language Learning), following the highly successful cases of Czech, English, Estonian, German, Italian, and Russian.

SkELL is a language learning tool that provides automatic summaries of corpus data.



The screenshot shows the SkELL web interface. At the top left is the SkELL logo. To its right is a search bar with a magnifying glass icon. Further right are navigation links: "Examples", "Word sketch", "Similar words", "More features", and "More languages". Below the search bar, the text "Sketch Engine for Language Learning" is displayed. A description follows: "Discover the English language through a billion-word collection of news, scientific papers, Wikipedia articles, fiction, web pages and blogs." Below this are three sections: "Examples: search for a word or a phrase and get short and meaningful example sentences for it.", "Word sketch is a list of words which occur frequently together with the searched word.", and "Similar words (not only synonyms) are words used in similar contexts visualized with a word cloud."



The ptSkELL project

The requisites for SkELL are:

- 1) A very **large corpus** with various genres? PtTenTen web corpus 3.8 bi words
- 2) **Sketch grammar** ✓ (Kuhn & Kosem, 2016)
- 3) **GDEX configuration** ✓ (Kosem, Koppel, Kuhn, Michelfeit, & Tiberius, 2018)

However, for language learning tools with **automatically-created** web corpus data, **further corpus cleaning up** is required.

NO offensive words nor sensitive content = NO PARSNIPs (Pork - Alcohol - Racism - Sex - Narcotics - Isms – Politics).



SkELL

The objective of task 5 (crowdfest) is to create a crowdsourcing project to help to clean the corpus, namely, take the whole corpus and find a way to get rid of inappropriate sentences

SkELL Exam

ravenously 0.08 hits per million

- 1 The long ride had made us **ravenously** hungry.
- 2 [REDACTED]
- 3 He used to eat it before **ravenously** for years.
- 4 Dogs, cats, and rats were **ravenously** devoured.
- 5 However, this state makes them **ravenously** hungry.
- 6 The nasty little critters love glucose and **ravenously** devour it .
- 7 She ate **ravenously** , suddenly even more hungry than she had imagined.
- 8 He would eat **ravenously** , and was particularly fond of snake meat.
- 9 Take nifedipine **ravenously** as noisy by your doctor.
- 10 Kris binges **ravenously** and falls asleep in soiled clothes.
- 11 She gave him the bacon and eggs and he set to **ravenously** .
- 12 She'd refused the airline food, and now found herself **ravenously** hungry.
- 13 I was **ravenously** hungry, and I severely needed to take a dump.
- 14 Zhu Bajie was very greedy, and could not survive without eating **ravenously** .

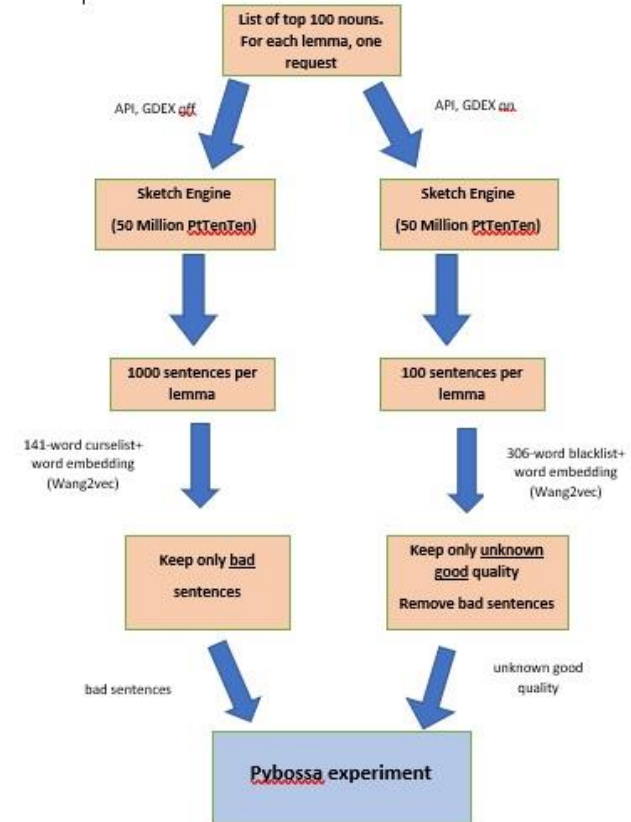
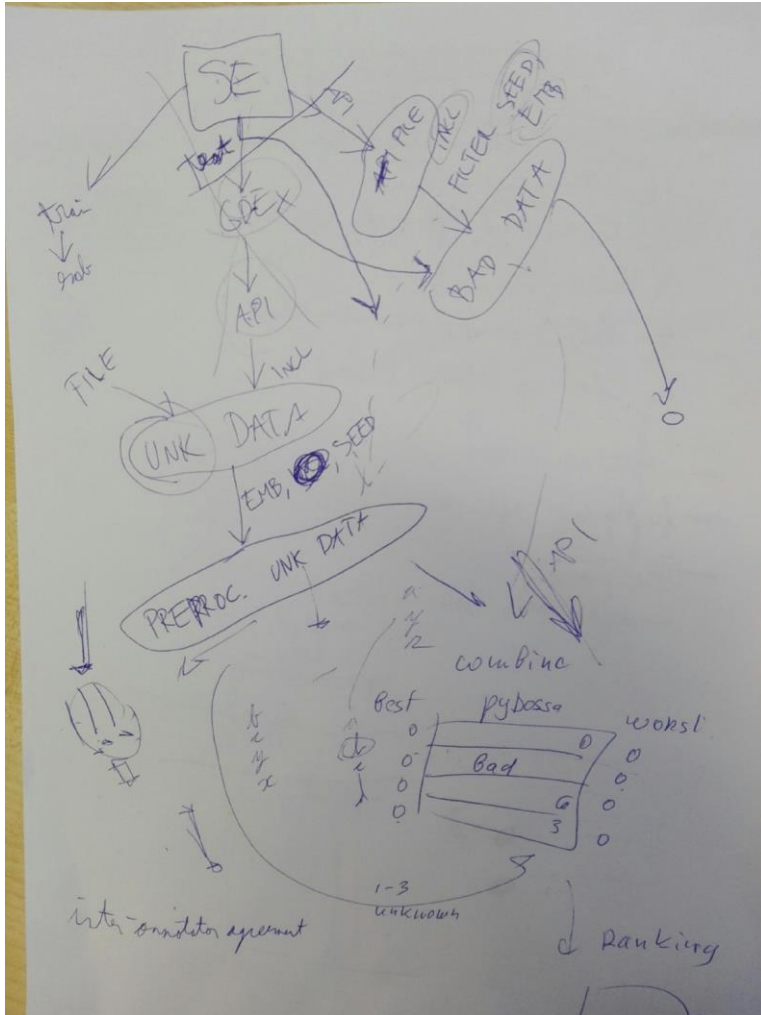


Research question

How can **web corpora** be cleaned for **language learning** purposes using **crowdsourcing**?



Crowdfest challenge: to develop a methodology



In two days...

- New blacklists
 - Long blacklist: (from Portuguese.gdex) - 306 words with cultural issues, leader names and countries (blacklist.txt)
 - Short curselist (from Portuguese.gdex) - 141 words - only really BAD words (curses), without the double meaning words and cultural issues (curselist.txt)
- API script with two options:
 - Option 1 - GDEX on - filtered sentences: 100 lemmas. Request 300 sentences from API. Remove duplicates. Keep 100 sentences. 100 sentences per lemma
 - Option 2 - GDEX off - unfiltered sentences: 100 lemmas. 1000 sentences per lemma
- Word embeddings to find synonyms of blacklist words:
 - Examined options: **SKIP-GRAM 300 dimensões** from NILC <http://www.nilc.icmc.usp.br/embeddings>, **Wang2Vec**
 - Based on experiments: **fasttext-skipgram 300-dim** model (from the same authors)
- Preliminary evaluation of extracted sentences
- Initial discussion about PYBOSSA task design



Sketch of crowdsourcing task

corpuscleanup-check: Contribua

Offensive

O diretor-presidente da Fundação Estadual de Pesquisa Agropecuária (Fepagro) , Marcos Palombini , disse em 2006 que o estado tem 1,4_milhão de hectares propensos a desertificar , e que naquele ano 25 % dessa área já estava comprometida . Ali é o domínio do pampa

Em 1963 , ainda criança , transferiu-se com a família para Tupãssi ; mesmo ano que ingressou no Seminário Menor de Toledo , onde fez o ensino fundamental

Carlos Staiger foi ainda um dos fundadores do Sindicato das Indústrias Metalúrgicas , Mecânicas e de Material Elétrico do Estado do Rio Grande do Sul (SINMETAL) , no qual ocupou diversos cargos , dentre os quais o de presidente entre os anos de 1983 e 1986

Andreas K. , de 37 anos , deu esta segunda feira uma entrevista ao jornal austríaco Kurier em que fala do crime que lhe levou toda a família

Expression

Save

Current task ID number: 748 .

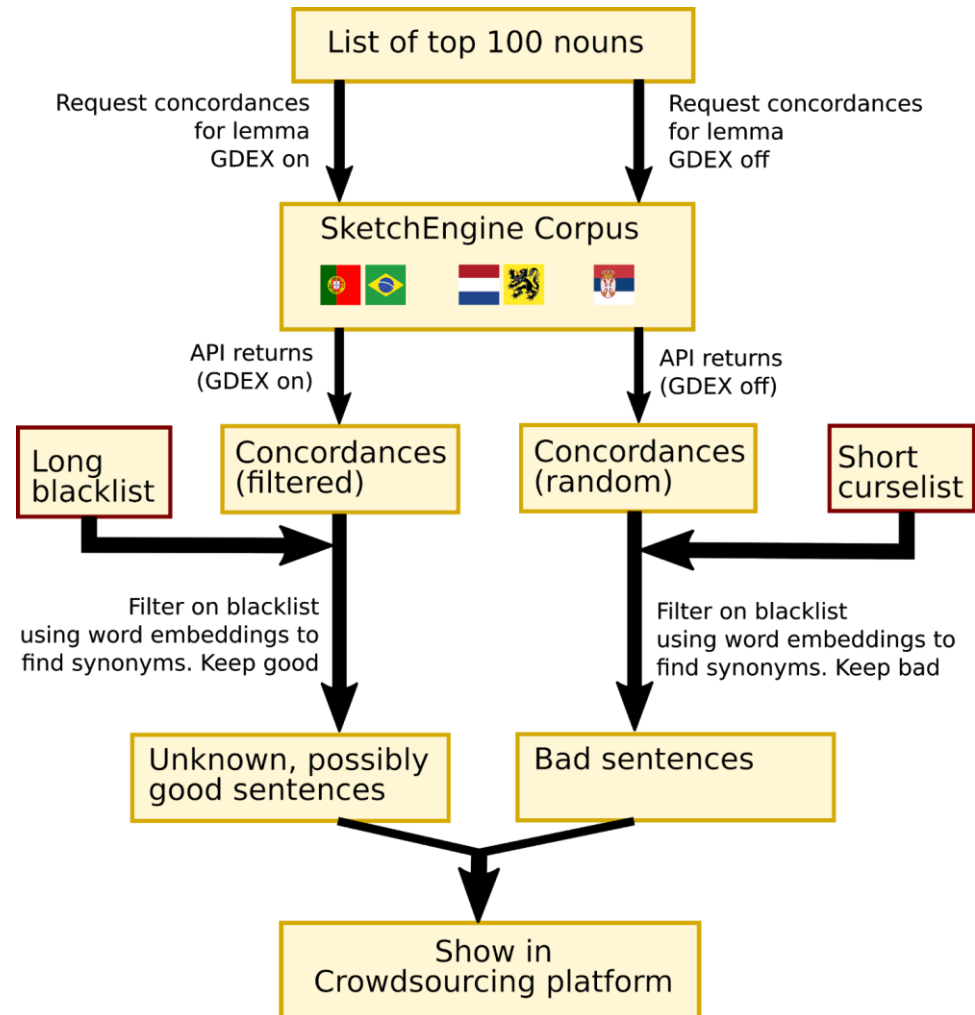
You have solved 0 task(s) out of a total of 1 . You are expected to solve 82 .

You can fill in the [feedback questionnaire](#) to describe how you made your decisions.



Since the crowdfest

- API script refactoring and extension
 - Multiple languages
 - Filtering no-GDEX results by GDEX length
- Data preprocessing for Serbian and Dutch
- One poster accepted – enetCollect
- Two conference papers submitted (Eurocall and eLex)



What's next?

- Two new team members (one new language – Slovene)
- 16th March meeting: crowdsourcing project design
- STSM?



References and Resources

- Pybossa: <http://pybossa.com/>
- SkELL: <https://skell.sketchengine.co.uk/run.cgi/skell>
- Kuhn, T.Z.; Kosem, Iztok (2016) Devising a Sketch Grammar for Academic Portuguese. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, v.4, p.124 - 161.
- Kosem, Iztok; Koppel, Kristina; Kuhn, Tanara Zingano; Michelfeit, Jan; Tiberius, Carole. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, advance article, p.1-19. DOI: 10.1093/ijl/ecy014. ISSN1477-4577.
- SKIP-GRAM 300 dimensões: <http://www.nilc.icmc.usp.br/embeddings>
- fasttext-skipgram 300-dim model:
http://143.107.183.175:22980/download.php?file=embeddings/fasttext/skip_s300.zip
- Hartmann et al. (2016) Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks.
<https://arxiv.org/pdf/1708.06025.pdf>



Thank you.
