Aranea:
Yet Another Family of (Comparable) Web Corpora

Vladimír Benko[1,2]

[1] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, SK-81101 Bratislava, Slovakia
[2] Comenius University in Bratislava, UNESCO Chair in Translation Studies
Šoltésovej 4, SK-81334 Bratislava, Slovakia
vladob@juls.savba.sk
http://www.juls.savba.sk/~vladob

Abstract. Our paper deals with an on-going Project in the framework of which, by means of open-source and free tools, a family of web corpora is being created that would (to a large extend) deserve the designation of being "comparable". A summary of results after the first stage of the Project is given, and experiences with the tools are commented.

Keywords: web-based corpora, web data filtration and deduplication, universal PoS tagset, compatible sketch grammar

1 Introduction

In spring 2013, a new web crawler, *SpiderLing*, has been released. This tool was the last missing stone in the mosaic of open-source and free tools for effective creation and annotation of web corpora. Thanks (mostly) to Computational Linguistics Departments at the Masaryk University in Brno and the University of Stuttgart, this mosaic now contains the following elements:

- *SpiderLing* [17] (a specialized crawler for downloading textual data from the web)
- *chared* (Python module for web page encoding detection, taking into consideration the expected language of the web page)
- *trigrams* (Python module for web page language detection)
- *jusText* [13] (utility for boilerplate removal)
- *Onion* [13] (deduplication utility based on n-grams)
- *Tree Tagger* [16] (tokenization and PoS tagging tool with parameter files for many languages
- *NoSketch Engine* [14] (corpus manager)

In our Project, by means of all the tools mentioned, we decided to create a family of web corpora that would (to a large extend) deserve the designation of being "comparable", i.e. the data would be downloaded at (approximately) the same time, they would contain similar (web-specific) composition of text types, genres and registers, would be of the same size, and would be available at one place via the unified access mechanism. The project can be described as Slovak-centric, as it

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 247–254, 2014.

should (in the first phase) cover the languages used and/or taught in Slovakia and the neighbouring countries, i.e. Slovak, Czech, German, Hungarian, Polish, Ukrainian and English, French, Spanish, Italian and Russian.

2 Corpus Design Decisions

Why we need other corpora? Besides our interest in testing the new corpus-building tools, the motive for starting our Project was the lack of suitable corpora that could be used by students of foreign languages and translation studies at our University. The existing corpora presented in [2] and [15] do not cover all languages needed. As for corpora described in [8] and hosted at the Sketch Engine web site,[3] they (1) are not available for downloading, (2) are typically too large for classroom use, and (3) have too different sketch grammars, which makes them difficult to use in a mixed-language classroom.

We expect that a family of corpora for several languages of equal size and built by standardized methodology can not only be used for teaching purposes, but also in linguistic research (contrastive studies) and in lexicography (both mono- and bilingual).

The names: For our corpora, we have decided to use "language-neutral" Latin names denoting the language of the texts and their size. The whole corpus family is called *Aranea*,[4] and the respective members bear the appropriate language name, e.g. *Araneum Anglicum*, *Araneum Germanicum*, *Araneum Russicum* for English, German and Russian, respectively, etc.

The sizes: Each corpus will exists in several editions, differing by their sizes. The basic medium-sized version, *Maius* ("greater"), will contain approximately 1.2 billion of tokens. This size is expected to contain at least 1 billion words, and can be reached relatively quickly for all participating languages. For the "large" ones with plenty of web data available it usually takes just one or two days to download the source data. The 10% random sample of *Maius*, called *Minus* ("smaller"), is to be used for teaching purposes. A 1% sample, *Minimum* ("minimal"), is not intended to be used directly by the end users, and is utilized in debugging of the processing pipeline and tuning the sketch grammars. And lastly, the largest *Maximum* ("maximal") edition will contain as much data as can be downloaded from the web for the particular language, and its size is mostly determined by the configuration of the server.

3 Crawling and Preprocessing

All source data acquisition is being performed by means of *SpiderLing*, a web crawler optimized for collecting textual data from the web. The system contains an integrated character encoding (*chared.py*) and language recognition (*trigrams.py*) module, as well as a tool for boilerplate removal (*jusText*).

The input seed URLs have initially been harvested by various methods. At present, the procedure has been standardized to consist of these steps:[5] (1) Take first two

[3] `http://www.sketchengine.co.uk`

[4] *Araneum* (pl. *aranea*, n.) is the Latin expression both for spider and (spider)web.

[5] The procedure has been partially inspired by [6].

paragraphs of the documents as follows: (a) Universal Declaration of the Human Rights, (b) Bible (John 1:1), (c) Wikipedia article for a concrete noun ("bicycle"), and (4) Wikipedia article for an abstract noun ("love"); (2) Tokenize and deduplicate the resulting wordlist, sort randomly; (3) Use the wordlist in several steps as seed for BootCAT [1], collect list of URLs (do not download the web pages themselves); (4) Deduplicate and filter the resulting URL list.

Using this method, we were quickly able to get several thousands of URLs that were subsequently used as seed for *SpiderLing*.

Several input parameters of the crawling process can (or must) be set by the user, most notably the language name, a file containing sample text in the respective language (to produce a model for language recognition), language similarity threshold (a value between 0 and 1 (default 0.5), number of parallel processes, and the crawling time.

In our processing, we usually crawled in 24-hour slots (the process could be later restarted) with all other values set to defaults. The only exception was crawling for Slovak and Czech, where we crawled in 7-day slots, as the process was much slower for these languages. The language similarity threshold had also to be changed in case of Slovak and Czech. As these languages are fairly similar, the trigram method did not seem to be able to distinguish between them sufficiently. We have therefore increased the threshold value to 0.65 (saving many "good" documents, and causing many "wrong" ones to pass the filter) and removed the unwanted texts by subsequent filtration based on character frequencies.[6]

Table 1 shows the share of eight most frequent top-level domains (TLDs) in documents for the respective languages (in percents).

Table 1. TLD distribution (in %).

de		en		es		fr		pl		ru		sk	
.de	71.34	.com	53.35	.com	45.52	.com	38.18	.pl	81.31	.ru	71.78	.sk	86.64
.com	10.55	.org	19.06	.es	20.11	.fr	33.41	.com	6.16	.com	10.95	.com	4.76
.at	5.26	.uk	6.67	.org	8.92	.org	9.86	.eu	4.69	.ua	6.42	.eu	3.99
.ch	3.78	.edu	5.10	.net	5.67	.net	5.45	.net	2.07	.org	2.97	.net	1.53
.net	3.34	.net	3.57	.ar	5.32	.ca	4.99	.info	1.80	.net	2.92	.cz	1.42
.org	2.76	.au	2.31	.mx	3.19	.be	3.09	.org	1.61	.info	2.56	.org	0.88
.info	1.57	.ca	1.85	.cl	2.92	.ch	2.20	.biz	0.51	.by	1.12	.info	0.54
.eu	1.18	.gov	1.53	.info	0.94	.info	1.59	.sk	0.36	.su	1.10	.rs	0.06
other	0.22	*other*	6.54	*other*	7.40	*other*	1.21	*other*	1.49	*other*	0.18	*other*	0.17

Quite consistent with our expectation, the national TLDs prevail in all languages spoken predominantly in a single country, and the "other" item is really significant only for languages spoken in many countries (English and Spanish).

Filtration: Besides the standard cleanup provided by the *SpiderLing* itself, we made use of some filters originally developed for our older Slovak web corpus, most notably

[6] The idea is (conceptually) based on counting frequencies of graphemes present in Slovak ("ä", "ľ", "ô"), and Czech ("ě", "ř", "ů") only, respectively.

to normalize representation of white space and special graphic characters, and to remove documents with misinterpreted encoding and/or having non-standard distribution of punctuation and uppercase characters (two few punctuation and/or too many uppercase chars usually mean that a page does not contain a "discursive" text). We also performed segmentation of the text on sentence boundaries by means of a rather rudimentary procedure (this segmentation was later used in deduplication).

Table 2 shows some statistics on the downloaded and preprocessed web data.

Table 2. Data downloaded, filtered and normalized.

	Domains	Docs	Tokens	Docs per domain	Tokens per doc
de	80,722	2,332,921	1,200,000,087	28.9	514.4
en	23,968	1,163,007	1,200,048,075	48.5	1031.8
es	22,343	1,439,567	1,049,739,252	64.4	729.2
fr	48,398	1,780,315	1,233,336,202	36.8	692.8
pl	58,338	1,783,411	1,110,120,825	30.6	622.5
ru	37,200	1,034,734	1,216,800,424	27.8	1176.0
sk	33,037	1,724,512	1,200,003,757	52.2	695.9

Deduplication: The whole procedure (described in more detail in our recent paper [3]) will finally consists of three stages. The first stage detects the near-duplicate documents by means of the Onion utility (similarity threshold 0.95), and the duplicate documents are deleted. The second stage deduplicates the remaining text at the paragraph level using the same procedure and settings. The tokens of the duplicate paragraphs, however, are not deleted but rather they are marked to make them "invisible" during corpus searches, while they can be displayed as context at the boundary of non-duplicate and duplicate text. In the last stage, we make use of our own tool based on the fingerprint method (with ignoring punctuation, special graphics characters and digits) to deduplicate the text at the sentence level. The tokens of duplicate sentences are marked similarly to the previous stage. This last step can "clean up" the duplicities among the short segments that fail to be detected as duplicates by Onion [13].

At present, only the fingerprint sentence deduplication has been used, and the whole procedure was postponed to later stages (to produce the upgraded version of our corpora). The results of the process can be seen in Table 3.

4 Linguistic Annotation

For all languages covered by parametric/dictionary files of *Tree Tagger* [16], this tagger has been used to annotate the respective corpora. For Polish, the *TaKIPI* [12], and for Czech, the *Morče* [7] taggers were used, respectively. The question of tools for PoS tagging of Hungarian and Ukrainian has not been resolved yet.

To simplify the creation of compatible sketch grammars, all native tagsets are mapped into the *Araneum Universal Tagset* (AUT) [4] (partially inspired by the Google

Table 3. Segmentation and deduplication.

	Sentences	Sentences per doc	Tokens per sentence	% of tokens removed
de	71,964,893	30.8	16.7	29.61
en	56,922,473	48.9	21.1	24.93
es	43,301,352	30.1	24.2	26.69
fr	54,650,594	30.7	22.6	26.57
pl	67,992,427	38.1	16.3	35.98
ru	69,180,355	66.9	17.6	21.10
sk	68,380,608	39.7	17.5	47.16

Universal PoS Tagset [11]) creating a secondary layer of morphosyntactic annotation. The AUT PoS tags are shown in Table 4.

Table 4. Araneum Universal Tagset.

aTag	PoS	aTag	PoS
Dt	determiner/article	Ij	interjection
Nn	noun	Pt	particle
Aj	adjective	Ab	abbreviation/acronym
Pn	pronoun	Sy	symbol
Nm	numeral	Nb	number
Vb	verb	Xx	other (content word)
Av	adverb	Xy	other other (function word)
Pp	preposition/postposition	Yy	unknown/alien/foreign
Cj	conjunction	Zz	punctuation

Besides the traditional 11 word classes, AUT contains 7 more items to accommodate information provided by the individual native tagsets (that is being merged into single a tag by the Google Universal PoS Tagset). Table 5 shows the share (in percents) of the respective word classes in seven Aranea corpora:

We can see that the numbers for the respective languages are surprisingly similar, with the exception of the "other" value Slovak, caused by some peculiarities of the SNK tagset.[7]

The subsequent filtration fixes some known tagger issues for the respective languages, namely the misassigned tags for several punctuation and special graphic characters (that are often tagged as nouns, verbs, or adjectives). For some languages, an additional tag with masked subcategories for gender and number is created, that can be later used by some rules within the respective sketch grammars.

[7] The SNK [5] tagset assigns a word class of its own for reflexive formants ("sa", "si") and for participles, with both having fairly high frequencies in the corpus.

Table 5. PoS distribution (in %).

	de	en	es	fr	pl	ru	sk
Dt	9.17	9.31	10.17	10.58	0.00	0.00	0.00
Nn	24.48	26.19	24.05	23.06	28.28	27.48	27.36
Aj	8.30	6.95	6.16	6.30	12.39	8.45	8.76
Pn	8.48	5.56	3.64	7.68	1.52	9.27	6.99
Nm	2.19	2.02	2.96	2.32	0.60	2.69	1.04
Vb	12.02	15.15	14.09	12.84	15.46	11.32	11.87
Av	5.21	4.83	2.97	4.86	2.10	3.64	2.32
Pp	9.06	10.44	12.72	15.43	10.20	9.22	9.23
Cj	5.12	3.42	7.72	4.18	5.87	6.18	5.85
Ij	0.02	0.05	0.00	0.06	0.00	0.07	0.04
Pt	1.71	0.34	3.11	0.00	5.32	2.68	2.62
Zz	13.64	12.50	11.82	12.11	5.02	18.99	14.55
other	0.61	3.24	0.58	0.57	3.22	0.00	9.38

5 Corpus Access

The standard environment for users to access the corpus data is the open-source *NoSketch Engine* developed at the Faculty of Informatics of the Masaryk University in Brno [14]. It is a mature, stable and user-friendly corpus manager offering all traditional concordancing- and wordlist-related search and display functions with queries based on wordform, lemma or PoS tag with optional use of regular expressions and the powerful Corpus Query Language (CQL). For users having an account at the Sketch Engine site, the installed versions of the *Aranea* corpora with compatible sketch grammars offer full capabilities of that system [10]. The source versions of the corpus data can be made available for download (for research and educational purposes). Note, however, that the copyright status of the data is not clear and users from countries where this might cause legal problems will have to solve this issue themselves.

6 The Sketch Grammar

For all corpora, compatible sketch grammars have been written. Their main idea is having an equal number of gramrels (and word sketch tables displayed) for all word classes across all languages. The principles and main design decisions behind creation of compatible sketch grammars are discussed in our work [4]. The Appendix contains an example of compatible word sketches generated from two *Aranea* corpora.

7 Current State of the Project

At the time of writing this Paper (May 2014), the basic medium-sized *Maius* (as well as the smaller *Minus*) *Aranea* editions for seven languages (Russian, French, German, Spanish, Polish, English, and Slovak) have been created, and compatible

sketch grammars have been written for all of them. For Slovak, the *Araneum Slovacum Maximum* (cca. 3 billion tokens) has also been compiled. Data for the Czech *Araneum Bohemicum* have been downloaded and filtered, and is being tagged at present. The downloading of data for the remaining languages of the "inner circle" (Hungarian, Ukrainian and Italian) will follow soon and the first stage of the Project is expected to complete by the end of 2014.

8 Conclusion and Further Work

The *Aranea* project has showed that by using the available open-source and free tools, billion-token web corpora can be created with minimal additional programming. After our processing pipeline has been tuned, a corpus for a new language (provided that a PoS tagger is available), including creation of a new sketch grammar, can typically be produced in some two weeks.

Our further activities are expected to follow several tracks. Firstly, based on the feedback from the users of our corpora, we would like to improve the data (filtration, tokenization, better deduplication, and tagging) of the existing corpora, and, where possible, to provide for alternative layer(s) of annotation, e.g. by using different taggers. Secondly, we want to include more languages into our *Aranea* corpus family, at least those taught as foreign languages at the Slovak universities (provided that suitable taggers exist for them). And lastly, we plan to compare the Aranea corpora among themselves and with other available web-based corpora for matching languages by means of methodology described in [9], and try to establish the degree of their mutual "comparability".

Acknowledgements The presented results were partially obtained under the VEGA Grant Agency Project No. 2/0015/14 (2014–2016).

References

1. Baroni, B., Bernardini, S.: BootCaT: Bootstrapping corpora and terms from the web. In: Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisbon (2004)
2. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43(3), pp. 209–226 (2009)
3. Benko, V: Data Deduplication in Slovak Corpora. In: Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning, pp. 27–39. RAM-Verlag, Lüdenscheid (2013)
4. Benko, V: Compatible Sketch Grammars for Comparable Corpora. In: Proc. XVI EURALEX Int. Congress, Bolzano (2014, in print)
5. Garabík, R., Šimková, M.: Slovak Morphosyntactic Tagset. Journal of Language Modelling, 0(1), 41–63 (2012)
6. Grefenstette, G.: Generating resources for the lexicography of under-resourced languages. Invited lecture at eLex 2013 Int. Conference, Tallinn (2013).
7. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Praha (2004)

8. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In: Proc. Int. Conf. on Corpus Linguistics, Lancaster (2013)
9. Kilgarriff, A.: Comparing Corpora. International Journal of Corpus Linguistics, 6(1), 97–133 (2001)
10. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proc. XI EURALEX Int. Congress, Lorient, pp. 105–116 (2004)
11. Petrov, S., Das, D., McDonald, R.: A Universal Part-of-Speech Tagset. In: Proc. 8th Int. Conf. on Language Resources and Evaluation, Istanbul (2012)
12. Piasecki, M.: Polish Tagger TaKIPI: Rule Based Construction and Optimisation. Task Quarterly, 11, 151–167 (2007)
13. Pomikálek, J.: Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno (2011)
14. Rychlý, P.: Manatee/Bonito – A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 65–70. Masaryk University, Brno (2007)
15. Schäfer, R., Bildhauer, F.: Web Corpus Construction. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. (2013)
16. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester (1994)
17. Suchomel V., Pomikálek J.: Efficient Web Crawling for Large Text Corpora. In: 7th Web as Corpus Workshop (WAC-7), Lyon, France (2012)

Appendix

Word sketches (collocation profiles) for the verb “drink” (“boire”) generated by Sketch Engine from *Araneum Anglicum Maius* and *Araneum Francogallicum Maius*. The gramrel (table) names denote collocational relationships. The “X” symbol stand for the keyword, i.e. the lemma the word sketch is made for. The left-hand and right-hand collocates are indicated by the respective PoS abbreviations. The “Y” stands for collocates of any PoS, and the “Z” indicates a collocate of PoS not covered by the “explicit” rules (i.e. a “catch all” rule).

Sketch Engine

drink *(verb)* Alternative PoS: non-verb (34500)

Araneum Anglicum Maius (En Web 1.2.01) 1,20 G freq = 42609 (35.5 per million)

X/Y , X/Y	5,596	-0.1
smoke	149	4.74
thirst	8	4.3
eat	709	3.63
party	8	3.15
gamble	10	2.92
breathe	31	2.05
dance	25	1.97
proclaim	23	1.9
sleep	42	1.31
marry	41	1.23
taste	16	0.83
shop	8	0.67
laugh	13	0.3
relax	13	0.22

X/Y Cj X/Y	10,312	-0.1
smoke	295	5.69
eat	2,951	5.68
carouse	15	5.39
party	43	5.37
drug	29	4.87
bathe	12	3.98
gamble	22	3.91
bath	31	3.72
dance	65	3.32
feast	11	3.29
snack	8	3.11
dine	21	2.86
socialize	12	2.6
drive	338	2.4

Y X	16,949	-0.0
binge	31	4.65
tritium	12	3.95
habitually	10	3.45
arsenic	13	3.38
quit	45	2.65
litre	11	2.65
contamination	18	2.58
Benefits	12	2.42
happily	18	2.4
contaminate	10	2.11
abstain	8	2.08
seldom	11	1.96
rarely	31	1.85
sport	71	1.66

X Y	27,833	-0.1
alcohol	805	5.95
Kool-Aid	51	5.75
beer	597	5.66
responsibly	83	5.64
tea	459	5.4
bottled	69	5.39
soda	123	5.37
coffee	591	5.28
alcoholic	121	5.27
champagne	76	5.11
wine	626	5.02
Kool	26	4.7
Jiaogulan	22	4.56
excessively	38	4.54

Nn X	16,383	-0.0
binge	33	4.77
Ye	34	4.6
tritium	17	4.48
drinker	21	4.0
arsenic	16	3.7
fluoride	15	3.67
gall	9	3.25
litre	16	3.2
alcoholic	25	3.16
wine	158	3.06
beer	90	2.97
cup	112	2.95
tea	80	2.92
beverage	30	2.9

X Nn	33,413	-0.1
tea	991	6.49
beer	1,032	6.43
alcohol	1,040	6.3
soda	237	6.23
Kool-Aid	83	6.21
coffee	1,095	6.16
smoothie	148	6.07
wine	1,129	5.86
juice	418	5.75
beverage	238	5.71
cup	755	5.65
milk	621	5.58
champagne	103	5.44
water	4,558	5.27

Aj X	4,639	-0.0
thirsty	37	5.56
underage	12	3.88
unfit	10	3.64
bottled	8	3.09
safe	200	2.71
unsafe	10	2.52
okay	21	2.13
clean	50	1.79
pregnant	24	1.67
drunk	9	1.54
ready	86	1.52
pleasant	14	1.51
OK	10	1.5
sick	19	0.88

X Aj	9,473	-0.1
bottled	93	6.42
caffeinated	31	6.26
koolaid	18	5.89
alcoholic	141	5.77
iced	32	5.55
sugary	24	4.97
contaminated	51	4.97
alkaline	30	4.92
fizzy	12	4.88
thirsty	26	4.81
copious	23	4.72
spirituous	8	4.68
non-alcoholic	9	4.35
herbal	49	4.33

Vb X/X Vb	44,353	-0.0
thirst	28	4.14
carbonate	25	4.02
hydrate	21	3.52
intoxicate	27	3.4
water	37	3.39
contaminate	33	3.39
milk	23	3.37
poison	30	3.35
fluoridate	14	3.26
purify	28	2.99
flush	28	2.97
boil	52	2.97
decaffeinate	10	2.82
vinegar	10	2.81

Av X/X Av	18,106	-0.1
responsibly	92	6.07
excessively	47	5.14
moderately	47	4.91
heavily	205	4.49
greedily	12	4.22
habitually	15	3.98
unworthily	9	3.96
sensibly	13	3.89
regularly	109	3.16
too	913	2.92
eagerly	14	2.85
freely	45	2.82
abundantly	9	2.79
anymore	46	2.77

Z X	32,200	-0.1
whoever	17	2.33
who	1,516	1.69
I	3,747	1.56
he	1,594	1.47
she	550	1.35
you	3,115	1.24
they	1,368	1.04
&	157	0.92
we	1,356	0.91
those	280	0.78
him	286	0.7
TO	8	0.68
What	83	0.55
to	8,542	0.54

X Z	25,135	-0.1
eight	80	2.01
himself	113	1.61
every	273	1.52
8	120	1.02
some	423	0.66
herself	17	0.63
six	61	0.61
myself	47	0.6
it	2,103	0.58
themselves	60	0.42
half	62	0.33
neither	11	0.28
whatever	20	0.24
2	186	0.22

Sketch Engine

boire **Araneum Francogallicum Maius (Fr Web 1.2.02) 1,23 G freq = 48230** (39.1 per million)

X/Y , X/Y	7,445	-0.1
fumer	311	5.21
droguer	25	4.91
manger	864	4.48
papoter	9	3.36
pisser	10	3.05
danser	54	2.75
dormir	78	2.3
rigoler	16	2.25
respirer	20	1.67
chanter	49	1.56
pendre	21	1.46
discuter	54	1.14
cuisiner	10	0.94
laver	17	0.92

X/Y Cj X/Y	9,211	-0.2
manger	2,362	5.93
fumer	268	4.99
droguer	25	4.82
uriner	18	4.33
grignoter	28	4.33
festoyer	9	4.13
enivrer	16	3.89
déboire	16	3.85
trinquer	11	3.81
pisser	9	2.84
danser	51	2.66
baigner	20	2.1
rigoler	13	1.93
dormir	56	1.82

Y X	26,598	-0.1
last	83	6.36
ossature	95	5.56
reputation	21	4.51
Serial	20	4.46
yogourt	23	4.31
yaourt	49	4.28
chaufferie	22	4.2
Last	21	4.14
serial	18	4.07
honte	100	3.88
nover	16	3.84
navette	48	3.71
pout	11	3.55
bardage	12	3.42

X Y	31,428	-0.1
gorgée	335	7.67
bière	689	6.66
verre	1,653	6.42
thé	670	6.11
not	116	5.87
calice	85	5.86
tasse	265	5.79
café	995	5.78
alcool	552	5.56
champagne	172	5.49
tisane	71	5.45
coca	79	5.42
potion	77	5.09
pisse	39	4.99

Nn X	17,569	-0.1
last	83	6.82
ossature	78	5.49
truck	30	5.19
universal	37	5.02
reputation	21	5.02
yogourt	23	4.7
chaufferie	22	4.57
yaourt	52	4.53
Last	21	4.53
pout	11	4.05
navette	56	4.04
honte	102	3.96
bardage	13	3.94
Rosny	9	3.73

X Nn	31,427	-0.1
gorgée	647	8.62
bière	936	7.1
verre	2,210	6.83
least	108	6.66
thé	840	6.44
tasse	385	6.33
calice	109	6.22
alcool	845	6.18
café	1,310	6.17
champagne	263	6.1
coca	112	5.92
tisane	93	5.84
whisky	86	5.57
soda	71	5.5

Aj X	3,627	-0.1
Serial	20	6.6
serial	18	5.6
Urban	15	4.17
rosé	15	3.95
préférable	20	2.07
agréable	61	1.97
potable	9	1.21
mixte	13	1.2
facile	70	1.11
prêt	73	0.77
mini	9	0.54
idéal	30	0.33
chaud	28	0.15
incapable	9	0.14

X Aj	6,084	-0.1
cul-sec	16	6.3
also	19	5.85
gazeuse|gazeux	24	4.73
empoisonné	10	4.66
alcalin	9	4.15
rosé	16	3.94
tiède	25	3.74
chaud	285	3.49
sec	78	3.31
you	9	2.69
digestif	11	2.46
frais	154	2.46
potable	21	2.41
amer	11	2.26

Vb X/X Vb	40,042	-0.1
not	152	6.08
gorger	49	4.34
it	44	4.06
epervier	18	3.86
alcooliser	33	3.82
sucrer	67	3.55
hydrater	30	3.41
nover	16	3.39
assoiffer	21	3.25
Vera	15	3.2
is	39	3.15
granuler	14	3.15
be	28	3.13
désaltérer	13	3.02

Av X/X Av	19,221	-0.1
modérément	36	5.36
avidement	19	4.67
tranquillement	99	4.48
jeun	22	4.47
abondamment	35	4.28
goulûment	10	3.65
suffisamment	132	3.39
lentement	50	3.19
régulièrement	194	3.11
occasionnellement	9	2.91
excessivement	9	2.89
trop	765	2.76
beaucoup	872	2.67
raisonnablement	10	2.66

Z X	35,379	-0.1
Donne-moi	27	4.53
donne-moi	13	3.49
Pouvez-vous	16	3.3
Assurez-vous	16	3.3
Peut-on	16	2.75
quiconque	23	2.46
on	2,456	2.24
tu	505	2.19
RER	9	2.05
quel	349	2.04
je	3,335	1.75
la	232	1.27
quoi	121	1.26
il	3,500	1.23

X Z	27,144	-0.1
+	94	1.5
quelque	425	1.06
mon	579	1.05
ton	138	0.94
&	81	0.91
~	26	0.84
un	7,147	0.83
L	26	0.8
I	21	0.78
=	41	0.67
son	1,424	0.39
huit	17	0.23
toi	32	0.14
chaque	135	0.09