

Marek Blahuš, Miloš Jakubíček,
Katarína Petreková & Yingtian Wu

AUTOMATIC DRAFTING OF TERMINOLOGICAL DICTIONARIES WITH ONE-CLICK DICTIONARY FOR 34 LANGUAGES

Abstract We present an evaluation of a system for automatic drafting of terminological dictionaries using the One-Click Dictionary feature of Sketch Engine. We bootstrapped five small corpora in Arabic, Czech, English, German and Spanish and evaluated the performance of the methodology on these five corpora yielding five terminological dictionaries for five different terminological domains; the system currently supports 34 languages in total.

Keywords Automatic dictionary drafting; terminology; Sketch Engine; Lexonomy; One-Click Dictionary

1 Introduction

Recent advances in NLP technologies including the availability of large language models significantly improved the quality of automatically created dictionaries. While general language dictionaries, both mono- and bilingual, still require significant amount of human post-editing to achieve publishing quality, terminological dictionaries represent a target much easier to achieve: their entries are (or at least should be) typically monosemous as each term should principally correspond to a single concept. This removes what is frequently seen as the most difficult lexicographic task: establishing on a word's sense division as well as crafting of further entry components such as collocations, examples or translations in a sense-specific way.

In this paper we present a processing pipeline available for 34 languages (Afrikaans, Arabic, Bosnian, Bulgarian, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Japanese, Korean, Latvian, Lithuanian, Maori, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, Vietnamese) which allows users of the system to automatically create a dictionary draft from a reasonably large (typically just dozens of millions of words) domain corpus. The pipeline is integrated in the Sketch Engine corpus management system (Kilgarrieff et al., 2014a) and consists of the following four steps:

1. building a domain corpus from users' texts or using web search
2. finding terms in the domain corpus using a language-specific term grammar
3. drafting a dictionary using the One-Click Dictionary feature
4. exporting the dictionary into the Lexonomy dictionary editing system

2 Building a domain corpus

The source evidence for the dictionary is gathered from a domain corpus supplied either by the user or created using web search. The former scenario is typical for large corporate users that

possess vast amount of internal documents describing their field and are looking for fast ways to create their own internal glossary. The latter option focuses on individual users or academic terminology and is based on the WebBootCaT technology (Baroni et al., 2006) which allows users to quickly gather texts from internet based on search seed words.

In our experiment we focused on the latter option and created five corpora for five domains. Languages were chosen with different language families in mind, domains were chosen according to the respective evaluators' subject knowledge. Details of the corpora can be found in Table 1.

TABLE 1 Description of corpora created in the experiment

Language	Size (tokens)	Domain
Arabic	11,124,898	football
Czech	15,763,412	book publishing
English	12,008,065	photography
German	6,544,280	automotive
Spanish	11,261,104	astronomy

We also provide an overview of the seed words used when crawling the web for corpus data using WebBootCaT. In accordance with the WebBootCaT best practices, a few seed words were identified for each language at the beginning, out of words that are typical of the domain (i.e., not necessarily of high frequency, but rarely used outside of the domain).

These seed words were used as the tool's input for the first pass, after which the process was iterated several more times, each time with a new set of seed words automatically discovered in the corpus data collected in the previous pass. This enabled us to grow the corpora to the desired sizes while keeping their content aligned with the selected domain. Following are the full lists of the used seed words:

- Arabic (86 seeds): *FIFA, أشرف حكيمي, يوفنتوس, أرسنال, أرسنال, مدردي, أتلتيكو, الأهلي, الاتحاد, الاتحاد الآسيوي لكرة القدم, الاتحاد الآسيوي لكرة القدم, الأفريقي لكرة القدم, الأفريقي لكرة القدم, الدوري الإسباني, الدوري الإنجليزي, الدوري السعودي, الدوري المصري, الدوري الإيطالي, الدوري الإيطالي, الرجاء, الزمالك, الفيفا, المرمي, النصر, الهلال, الوداد, الوقت الإضافي, انتر ميلان, بارس سان جرمين, بارس سان جيرمان, بايرن ميونخ, برشلونه, بروسيا دورتموند, بطاقة الحمراء, بطاقة صفراء, بطولة كأس العالم, بروسيا دورتموند, تسجي هدف, تسديدة, تسلي, تشيلسي, تشيلسي, حارس المرمى, حارس المرمى, حجم الأصلي للصورة, حكيمة زياش, دوري أبطال أوروبا, دوري أبطال أوروبا, رأس, ربع النهائي, ركل, ركلة جزاء, ركلة حرة, ركلة ركنية, ركل حرة, رياض محرز, ريال مدريد, ساديو ماني, شوط أول, شوط ثاني, شوط ثاني الوقت الإضافي, كأس آسي, كأس الأمم الأفريقية, كأس العالم, كأس العالم للأندية لكرة القدم, كريستيانو رونالدو, كريم بنزيمة, كيليان مبابي, ليفربول, ليونيل ميسي, مانشستر سيتي, مانشستر يونايتد, مبادرة كرة القدم, مباراة كرة القدم, مباشرة حرة, محمد صلاح, منتخب المغرب, منطقة الجزاء, ميلان, ميونخ, نهاية الوقت, نيما, هلال*
- Czech (59 seeds): *akvizice rukopisů, antikva, autotyp, bohemikum, copyright, čtverčík, distributor, ediční plán, formát, fraktura, gramáž papíru, hřbet knihy, iniciála, inkunabule, ISBN, kalandr, kaligrafický, kapitálek, kapitálka, knihkupec, knihtisk, knižní blok, knižní trh, knižní vazba, korektura, kurzíva, litera, maloobchodní cena knihy, marketingové strategie nakladatelství,*

marketingový mix, minuskula, minuskule, nakladatel, patitůl, pečeť, písemnictví, pomlčka, produkce knižních titulů, propagace knihy, průměrná cena, průměrný náklad, prvotisk, předmluva, předsádka, signet, signum, spojovník, státní podpora knižního průmyslu, struktura knižního trhu, štoček, švabach, tiráž, tiskař, tištěná kniha, titul, trendy knižního trhu, typografický, vakát, verzálka

- English (194 seeds): *aperture, aperture priority, APS-C, artificial light, aspect ratio, astrophotography, autofocus, auto mode, backlighting, ball head, batch editing, black-and-white photography, blending, blue hour, blur, blurred background, blurring, bokeh, bracketing, brightness, built-in flash, bulb mode, camera, camera body, camera lifespan, card reader, cataloging, clone stamp, color photography, compact camera, composition, contrast, conversion, copyright, couple photography, cropping, data backup, data recording, depth of field, developer, digital camera, display, distracting background, DSLR, EXIF, exposure, exposure mode, exposure time, exposure triangle, exposure value, extender ring, external flash, family photo, filter, final photo, fixed focal length, fixed lens, fixer, focal length, focus point, foreground, freezing motion, glare, golden hour, gradation filter, gradient filter, graphics card, halogen bulb, hard drive, HDR, histogram, hyperfocal distance, chromatic aberration, image, image quality, image sensor, image stabilization, incandescent bulb, interface, interior, ISO, JPEG, landscape photography, lens, lens distortion, lens hood, light-sensitive, light source, live view, local editing, long exposure, macro lens, macro photography, magnetic filter, main subject, manual focus, manual mode, memory card, minimalist photography, mirrorless camera, monitor, monitor calibration, motion blur, mount, mount adapter, natural light, negative, neutral density filter, noise, noise reduction, open aperture, operating system, optical stabilization, outdoor photography, overexposed image, panning, panorama, parameter, perspective, photo atmosphere, photo background, photo download, photo export, photogenic, photograph, photographed scene, photographed subject, photography equipment, photography school, photo import, photo layers, photo management, photo processing, photo resolution, photo series, photo sorting, pixel, PNG, polarizing filter, portrait, preview, product photography, RAM, raw development, RAW format, raw photo, read speed, reflection, reflector, reportage photography, reproduction, retouching, retouching tool, RGB, ring light, self-timer, sensor, sensor sensitivity, shadow, sharpening, sharpness, shutter, shutter button, shutter priority, shutter release, shutter speed, smudging, softbox, softening light, software, sRGB, storage, street photography, studio, technical quality, telephoto lens, thematic photography, TIFF, time-lapse, tonality, tripod, tripod plate, underexposed image, USM, variable filter, video mode, viewfinder, vignetting, wedding photography, white balance, wide-angle lens, wildlife photography, zoom lens, 35mm film*
- German (37 seeds): *Abgasnorm, Akkufüllung, Assistenzsysteme, Außenspiegel, Direktschaltgetriebe, Einparkhilfe, Fahrgastraum, Fahrstil, Fahrzeugklasse, Familienauto, Federungskomfort, Fernlichtassistent, Feuchtigkeitssensor, Fünf-Liter-Marke, Geschwindigkeitsbegrenzer, Heckscheibe, Höchstgeschwindigkeit, Infotainmentsystem, Innenspiegel, Karosserie, Kofferraumvolumen, Kraftstoffverbrauch, Kurvenräuber, Limousine, Mittelkonsole, Motorabdeckung, Neuwagenpreis, Panoramadach, Plug-in-Hybrid, Reichweite, Rückfahrkamera, Serienausstattung, siebensitzig, Spurhalteassistent, Vierzylinder, Vorderrad, Winterreifen*
- Spanish (73 seeds): *agujero negro, agujero negro supermasivo, año luz, año luz de distancia, astronomía, astronómico, astrónomo, Betelgeuse, campo magnético, cinturón de asteroides, cometa, constelación, cuerpo celeste, cúmulo, cúmulo abierto, cúmulo de galaxias, cúmulo estelar, cúmulo globular, elemento más pesado, enano blanca, enano roja, energía oscura, esfera celeste, estrella, estrella cercana, estrella de neutrones, estrella masiva, exoplaneta, expansión del universo, explosión de supernova, formación estelar, fusión nuclear, galaxia, galaxia espiral, gigante roja, Hubble, interestelar, Júpiter, longitud de onda, luminosidad, luz de distancia, magnitud aparente, masa solar, materia oscura, NASA, nebulosa, nebulosa planetaria, neutrino, neutrón, NGC, nube de gas, nube de gas y polvo, objeto astronómico, planeta enano, planeta extrasolares, planeta terrestre, polvo interestelar, protoestrella, pulsar, radiación, rayo gamma, sistema binario, sistema*

estelar, sistema planetario, sistema solar, supergigante, supernova, telescopio, telescopio espacial, telescopio Hubble, universo observable, velocidad de la luz, velocidad de rotación

3 Term finding

Having domain corpus texts, by uploading them into Sketch Engine, each corpus gets automatically analyzed by a part-of-speech tagger and a lemmatizer. Such morphological annotation is used for finding candidate terms using a language specific term grammar – a morphosyntactic description of grammatically valid terms (typically noun phrases) in the form of corpus queries in the CQL language (Kilgarriff et al., 2014b).

The term grammars are developed to achieve high coverage on existing term bases such as IATE and so far there are term grammars for 34 languages, which is the present restriction on the languages supported. In One-Click Dictionary, the extracted terms are used as input (i.e., headwords) for further dictionary drafting. This is similar to how terminologists routinely use the same algorithm through a dedicated interface called OneClick Terms to extract terminology and then build term bases or glossaries out of it using specialized terminology management tools (Blahuš et al., 2023).

4 One-Click Dictionary

Finally, a functionality of Sketch Engine called One-Click Dictionary (Jakubíček et al., 2021) is used to create a dictionary draft automatically (with “one click”) from a corpus. The dictionary headwords are sourced from the most frequent single- and multi-word terms in the source corpus as identified by the respective term grammar. Further entry components are generated from the source corpora:

- word forms; based on the automatic lemmatization of the corpus
- example sentences; based on the GDEX functionality of Sketch Engine
- collocations; based on the word sketch functionality of Sketch Engine
- thesaurus; based on the thesaurus functionality of Sketch Engine
- images; sourced from royalty-free databases such as Wikimedia Commons and Pixabay
- definitions; based on a large language model which takes all the other entry components (and the domain specification as well, of course) as input for consideration, leading to very high accuracy of the generated definitions. The system allows the user to choose from dozens of models provided by the OpenRouter platform. In this experiment, we used the GPT-5 model released in August 2025.

Each of the steps that sources data from the corpus comes with a settings allowing the user to select the number of example sentences, collocations, thesaurus items etc., including selection criteria as available elsewhere in the Sketch Engine corpus management system. Figure 1 shows the components used in our experiment.

As explained earlier, no word sense induction or disambiguation is performed as we build upon the assumption that a terminological entry should always correspond only to a single sense for a given fixed domain.

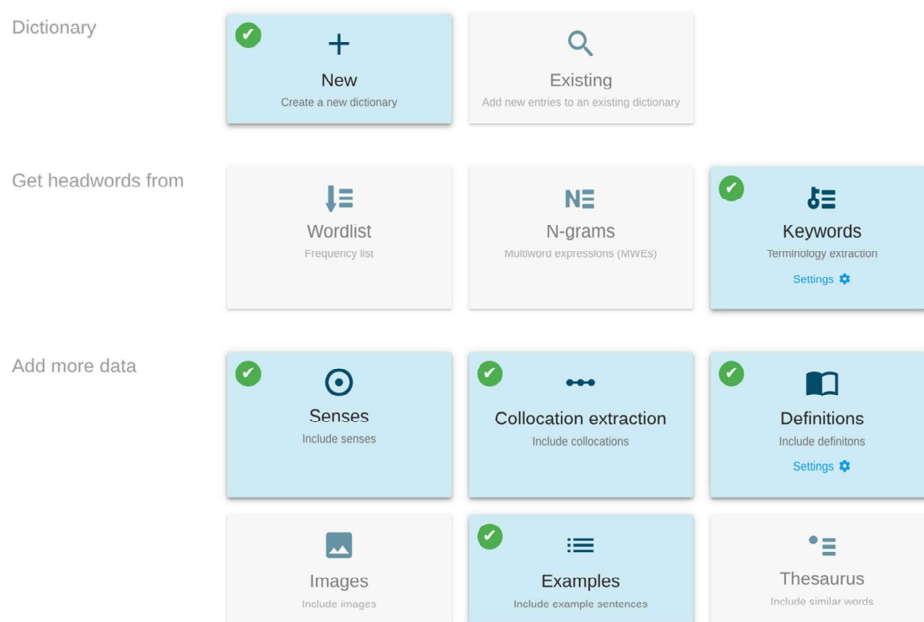


FIGURE 1 One-Click Dictionary user interface

The fully automatic dictionary drafting using One-Click Dictionary with the above settings takes less than 60 minutes for the first 1000 dictionary entries.

5 Lexonomy

The ready-made dictionary draft is made available in Lexonomy (Jakubíček et al., 2018), a web-based dictionary editor, where it can be reviewed, optionally post-edited (preserving access to the underlying corpus evidence in Sketch Engine) and published. In Figure 2 you can find a sample entry *exposure time* generated for the photography domain from the English corpus.

exposure time

[see concordance](#) CQL: `[term("(?i)exposure time")]`

[see word sketch](#)

Exposure time refers to the duration for which a camera's sensor or film is exposed to light, directly influencing the brightness and motion blur in an image.

sense defining collocations: exposure (noun), expose (verb), potentially (adverb), prolonged (adjective), negative (adjective)

► **modifiers of "exposure time"**

[see concordance](#)

CQL: `[swap(1, swap(1, ccoll(1, 2, swap(1, ws(2, 6650934638))) & ws(1, 31209926)) & lepos!="exposure-n" & lepos!="time-n")]`

maximum

[see concordance](#) CQL: `[ccoll(1, 2, swap(1, ws(2, 6650934638))) & ws(2, 3315689549)]`

- The maximum exposure time for Schedule Captures is 60 seconds.
- In Auto Mode, the maximum exposure time is 15 seconds.
- Many limit their maximum exposure times to the order of a minute.

► **verbs with "exposure time" as subject**

[see concordance](#)

FIGURE 2 Sample English entry from the photography corpus

6 Evaluation

We evaluated the automatically created dictionary drafts as for the plausibility of headwords and accuracy of examples and definitions. Each dictionary was assessed by a single reviewer who was a proficient speaker of the respective language and domain expert for the respective topic. Only a hundred top-ranked terms of each dictionary have been evaluated. Table 2 presents results of the evaluation for the respective languages.

TABLE 2 Evaluation results assessing the quality of the dictionary drafts as for headwords, examples and definitions. The percentages for examples and definitions are calculated from plausible headwords only. Use the ID to access the corpus in Sketch Engine (in the Select corpus page, or in a URL of type <https://app.sketchengine.eu/?corpname=preloaded/ID>) or the dictionary in Lexonomy (in the list of public dictionaries, or in a URL of type <https://www.lexonomy.eu/ID>)

Language	Headword plausibility	Definition accuracy	Example accuracy	Corpus and dictionary ID
Arabic	88.0 %	97.7 %	91.1 %	euralex2026_ocr_ar
Czech	84.0 %	90.5 %	57.9 %	euralex2026_ocr_cs
English	97.0 %	96.9 %	71.1 %	euralex2026_ocr_en
German	88.0 %	96.6 %	79.4 %	euralex2026_ocr_de
Spanish	94.0 %	94.7 %	73.4 %	euralex2026_ocr_es
Average	90.2 %	95.3 %	74.6 %	

A headword was considered plausible if it was a domain term for the respective domain. Minor deviations from the expected citation form were tolerated, such as incorrect suffixes or wrong letter case, and proper nouns were accepted as terms as long as they identified a specific named entity in the domain. Semantic overlaps between terms, such as synonymy or hyponymy (e.g. the Spanish *agujero negro* 'black hole' vs. *agujero negro supermasivo* 'supermassive black hole'), were disregarded during evaluation.

A definition or example was considered accurate if the reviewer finds it suitable for dictionary inclusion without further changes. While the system tries to come up with multiple senses (and therefore multiple definitions) for each headword, we found it sufficient to consider only the first one, supported by the expected monosemy of terms.

The number of generated examples depends on the number of identified collocations. In order to balance evaluation costs and representativeness, we assessed three examples for each headword: the first one, the last one, and one from the middle of the list.

Public access is given to all the corpora used (in Sketch Engine) and the respective drafted dictionaries (in Lexonomy).

7 Conclusions and future work

In this paper we presented a minimal framework that demonstrates rapid development of terminological dictionaries using domain seed words, corpus bootstrapping and automatic dictionary drafting. The framework could be extended to additional languages and domains in a straightforward way, providing stable benchmarking for the task of automatic dictionary drafting for terminological dictionaries.

References

- Baroni, M., Kilgarrieff, A., Pomikálek, J. & Rychlý, P. (2006). WebBootCaT. Instant domain-specific corpora to support human translators. In *Proceedings of the 11th annual conference of the European association for machine translation*.
- Blahuš, M., Jakubíček, M., Cukr, M., Kovář, V. & Suchomel, V. (2023). Development of evidence-based grammars for terminology extraction in OneClick Terms. In *Proceedings of the Electronic Lexicography in the 21st Century, Brno, Czech Republic* (pp. 27–29). Lexical Computing.
- Jakubíček, M., Kovář, V. & Rychlý, P. (2021). *Million-click dictionary: Tools and methods for automatic dictionary drafting and post-editing*. In Z. Gavriilidou, L. Mitits & S. Kiosses (Eds.), *EURALEX XIX* (pp. 65–67). SynMorPhoSe Lab, Democritus University of Thrace.
- Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-editing Lexicography with Lexonomy and Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *The XVIII EURALEX International Congress: Lexicography in Global Contexts. Book of Abstracts* (p. 65).
- Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014a). The Sketch Engine: ten years on. *Lexicography*, 1. <https://doi.org/10.1007/s40607-014-0009-9>
- Kilgarrieff, A., Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (2014b). Finding terms in corpora for many languages with the Sketch Engine. In S. Wintner, M. Tadić & B. Babych (Eds.), *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 53–56).

Contact information

Marek Blahuš
Lexical Computing, Brno, Czech Republic
marek.blahus@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing, Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Katarína Petreková
Lexical Computing, Brno, Czech Republic
Faculty of Arts, Masaryk University, Brno, Czech Republic
katarina.petrekova@sketchengine.eu

Yingtian Wu
Lexical Computing, Brno, Czech Republic
wyt15858810001@163.com