ELEXIS
LEX2 infrastructure

Miloš Jakubíček, Lexical Computing

February 18, 2019
ELEXIS observer event, Vienna

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

LEX2 infrastructure

- virtual access (= online only)
- single-sign-on (SSO) using eduGAIN network
- freely available for non-commercial usage across EU countries
- as of February 2019:
  - Sketch Engine
  - Lexonomy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

Sketch Engine

- corpus query system (CQS)
- platform for providing and sharing corpora
- platform for building corpora

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Sketch Engine: corpus search

- concordance
- wordlist of words/ngrams
- word sketch
- sketch diff
- distributional thesaurus
- terminology extraction
- trends detection

...and much more

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Sketch Engine: building corpora from web

- web crawling – Spiderling
- (partial) text deduplication – Onion
- boilerplate removal – Justext
- character encoding detection – Chared
- character normalization – Uninorm
- tokenization – Unitok

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Sketch Engine: language pipelines

- PoS tagging: 36 languages
- lemmatization: 32 languages
- word sketch grammar: 30 languages
- term grammar: 19 languages

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Sketch Engine: corpora

- corpora for 100+ languages
- 85 lgs over 10M
- 54 lgs over 100M
- 31 lgs over 1G
- 12 lgs over 10G
- 63 lgs with a tagged corpus
- 55 lgs with a lemmatised corpus

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Sketch Engine: corpora

- web corpora
- monitor corpora
- learner corpora
- parallel corpora
- spoken corpora
- ...

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Lexonomy

- simple web-based dictionary writing system
- interconnected with Sketch Engine (push & pull)
- designed for post-editing
- more to follow by Iztok Kosem and Michal Měchura

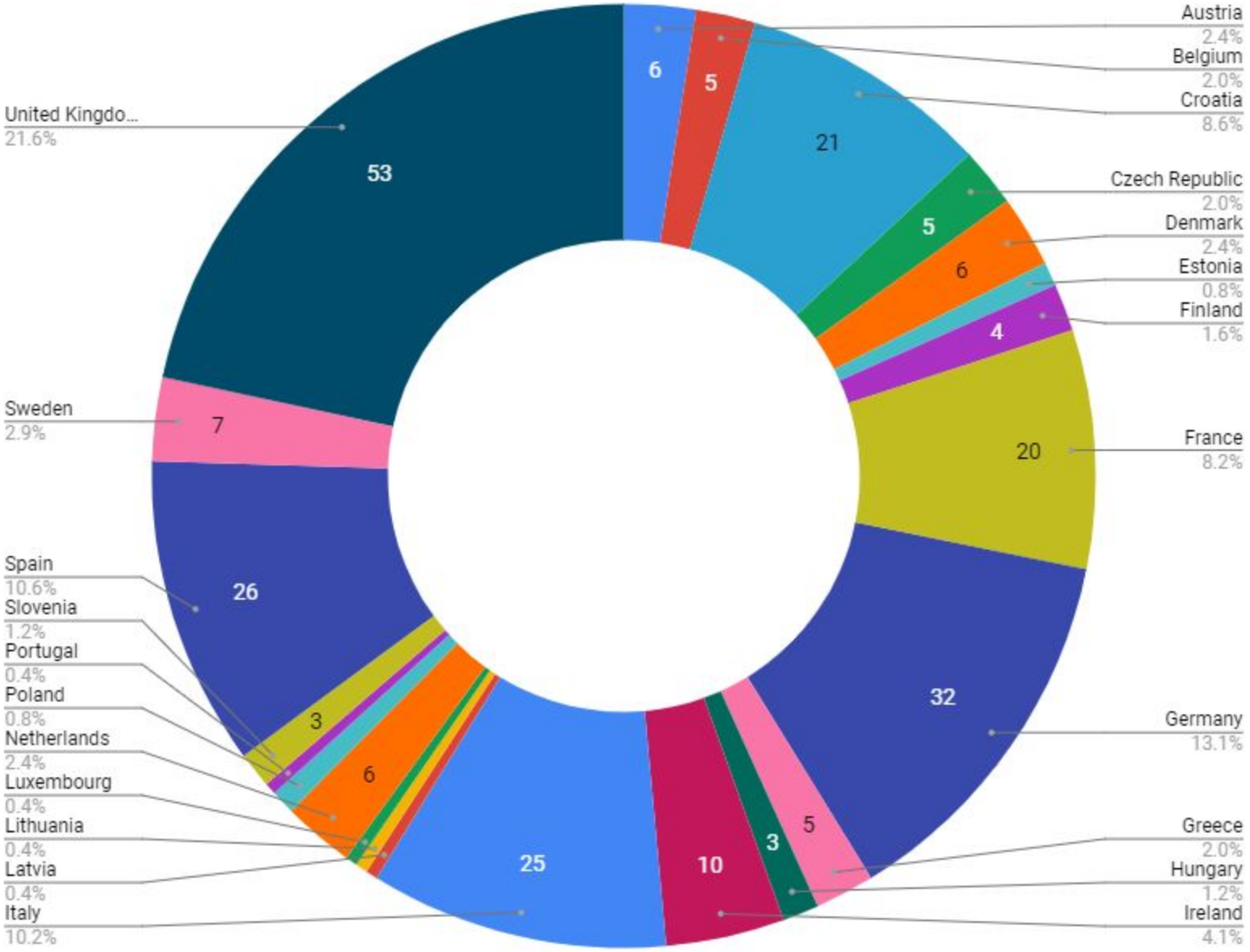
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Status of the LEX2 infrastructure

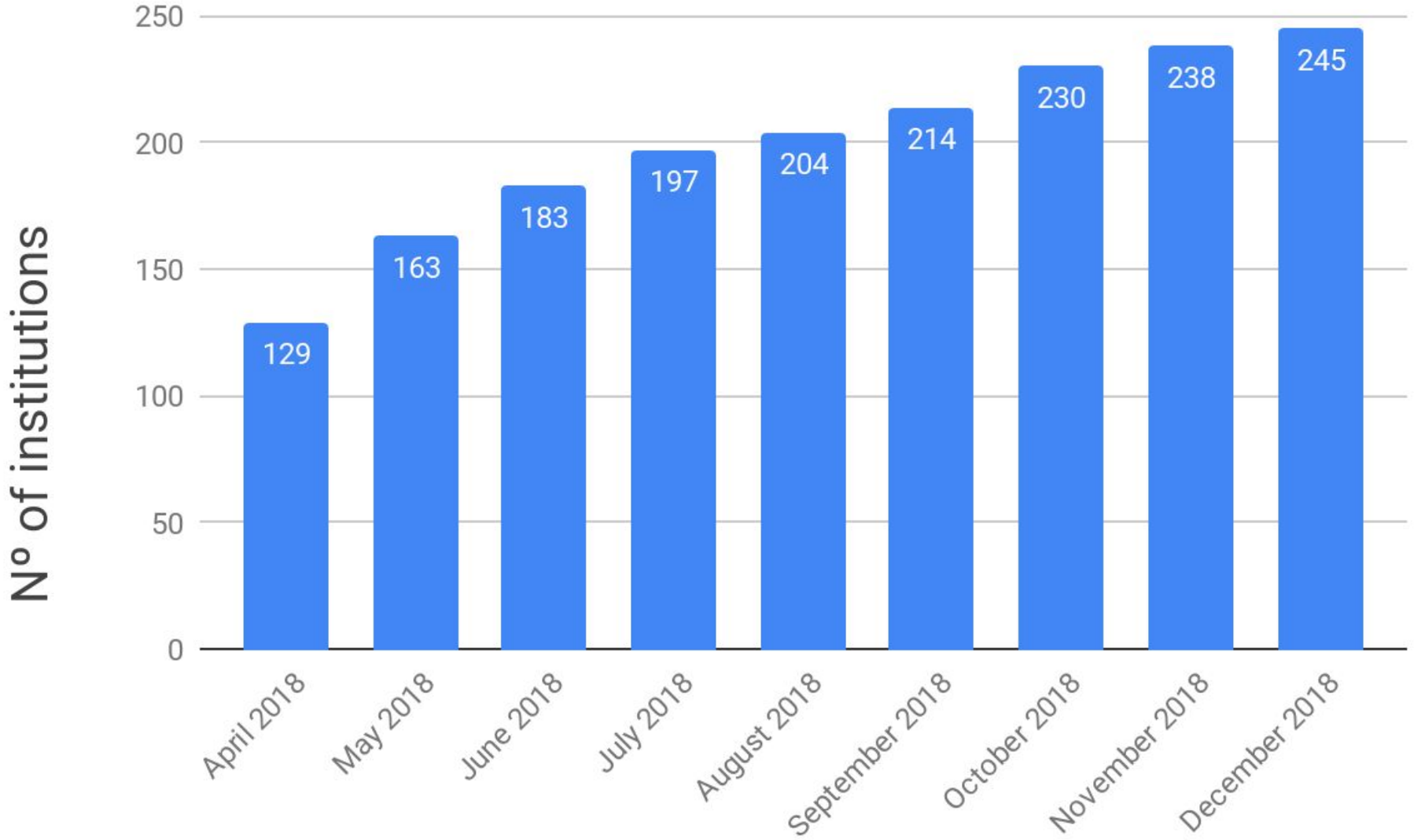
As of February 2019:

- 265 institutions joined the infrastructure from 23 EU countries
- over 10,000 users
- over 150,000 user-hours

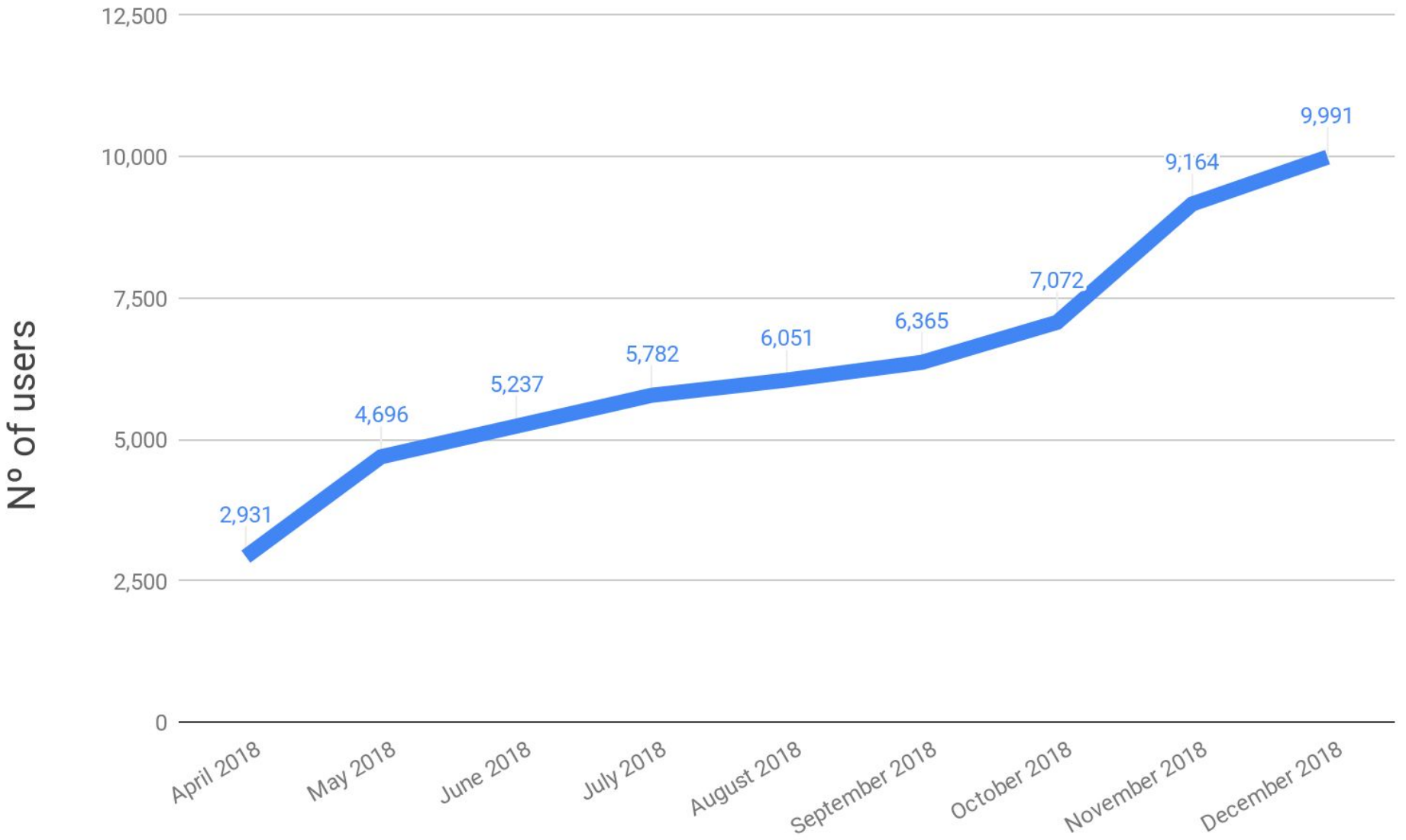
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



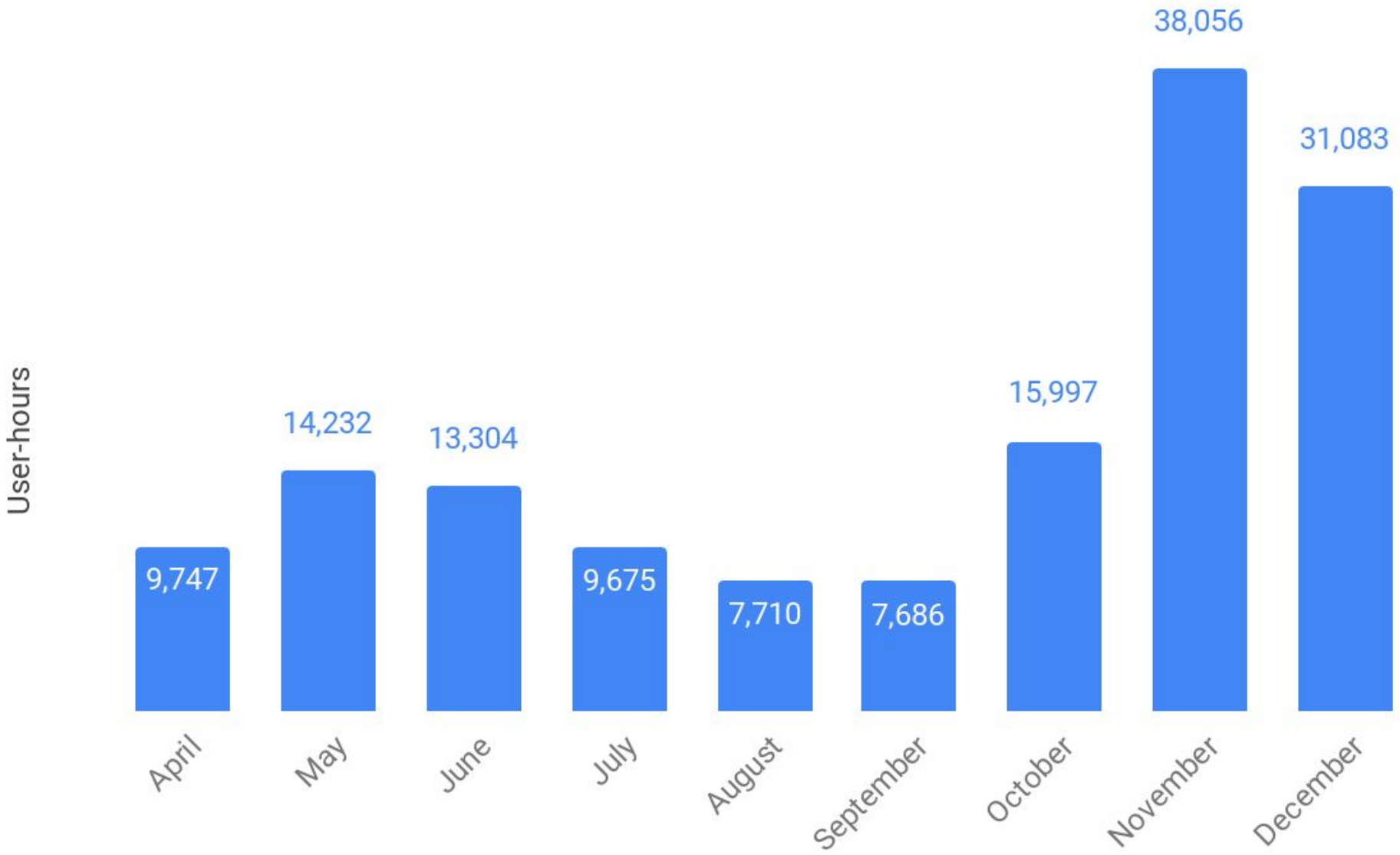
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Current and future development

Expanding the infrastructure (WP8)

- further tools and services
- e. g. Skema: pattern editor following a CPA approach (Hanks et al.)

Research and development (WP4)

- NLP for lexicography
- dictionary drafting and post-editing

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary drafting

- One-Click Dictionary function in Sketch Engine
- a dictionary draft automatically pushed to Lexonomy

- headwords: wordlist
- word senses: clustering through word sketches
- collocations: word sketches
- examples: GDEX
- definitions: GDEF
- labels: word sketch highlights
- translations: bidicts from parallel corpora

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary post-editing

- development of tools that support efficient post-editing
- development of methodology that accounts for post-editing
- teaching and training
- all included into Lexonomy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Conclusions

- LEX2 infrastructure ready and growing fast
- more tools and better tools to be included soon
- any questions ⇒ talk to us at the booth!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.