(Doplnění minulých hodin)

PLIN035: Počítačová lexikografie

František Kovařík

Opáčko

Citlivá témata aka PARSNIP

P – Politics 👮 🗳️ 🗣️ 📣 👔 🌎 🇺🇸 🇷🇺 🇨🇳 🇰🇵 🇮🇱

A – Alcohol 🍺 🍾 🍻 🥂 🍸 🍶 🍷

R – Religion ⛪ 🕌 🛕 ⛩️ 🕋 ⛲ 🙏 🙇 🤦 😇 ✝️ ☦️ ✡️ ☪️ 🛐 ☯️ ☸️ ⚛️ 🕉️ 🕎

S – Sex 💑 💕 😉 😗 🔞 👙 💋 🏩 ♀️♂️

N – Narcotics 💉 💊 ⚗️ 🚬 🌿 🍄

I – "-isms" (e.g. communism, atheism…) 🧠 💭 🗽 ⚛️ 🏳️‍🌈 📚 👻 📜

P – Pork (i.e. cultural taboos) 🐷 🥩 🐄 🌱 禁

– slova vysvětlovat, ale nehodnotit, neprojektovat svoje (kulturní) očekávání (jenže… nakolik jsme si jich *u sebe* vědomi?)

Příklady ve slovníku

Diachronní:

- doklad užití;

- změny v jazyce.

Příklady ve slovníku

Lexikální databáze:

ilustrace užití

– syntaktické vzorce, kolokace.

Příklady ve slovníku

Ve výkladovém slovníku:

doplnění definice;

typické užití.

Příklady z korpusu

Upravovat?

Neupravovat?

Neupravovat…

… pokud jsou v souladu s potřebami uživatele.

+ Dostatečně velký korpus.

Upravovat

Ilustrace vzorců a ko-textu.

To edit or not to edit: some test cases for the word *allegation*

- *How would you answer the allegation that it is unduly restrictive,…*
  1. *answer* is not the most typical verb collocate (*deny, refute, counter* are far more frequent - see Word Sketch)
  2. we don't know what *it* refers to: pronouns can cause problems
- Not a good basis for a dictionary example

To edit or not to edit

- *The allegations have been strongly denied by the government, which insists that ~~all agreed water quality standards are publicly available, and that all~~ drinking water is safe.*
- Good example, but too long: just delete the part shown
- The key content (*allegations were strongly denied*) remains intact

To edit or not to edit

- *He has proof to back up allegations of 'jobs for the boys' in Monklands District Council.*
- *back up* + *allegations*: a good combination, but
  - users may not know the phrase *jobs for the boys*
  - *Monklands District Council* could be distracting (and may raise legal issues)
  - Better something like: *He has proof to back up his allegations of corruption in the council.*
  - Again, this preserves the core of the example.

Dobrý příklad

Zachycuje, co je typické.

Je informativní.

Je srozumitelný.

Zachycuje, co je typické

Typická syntax, kontext, kolokace.

Nedobré příklady

„dělat si starosti“

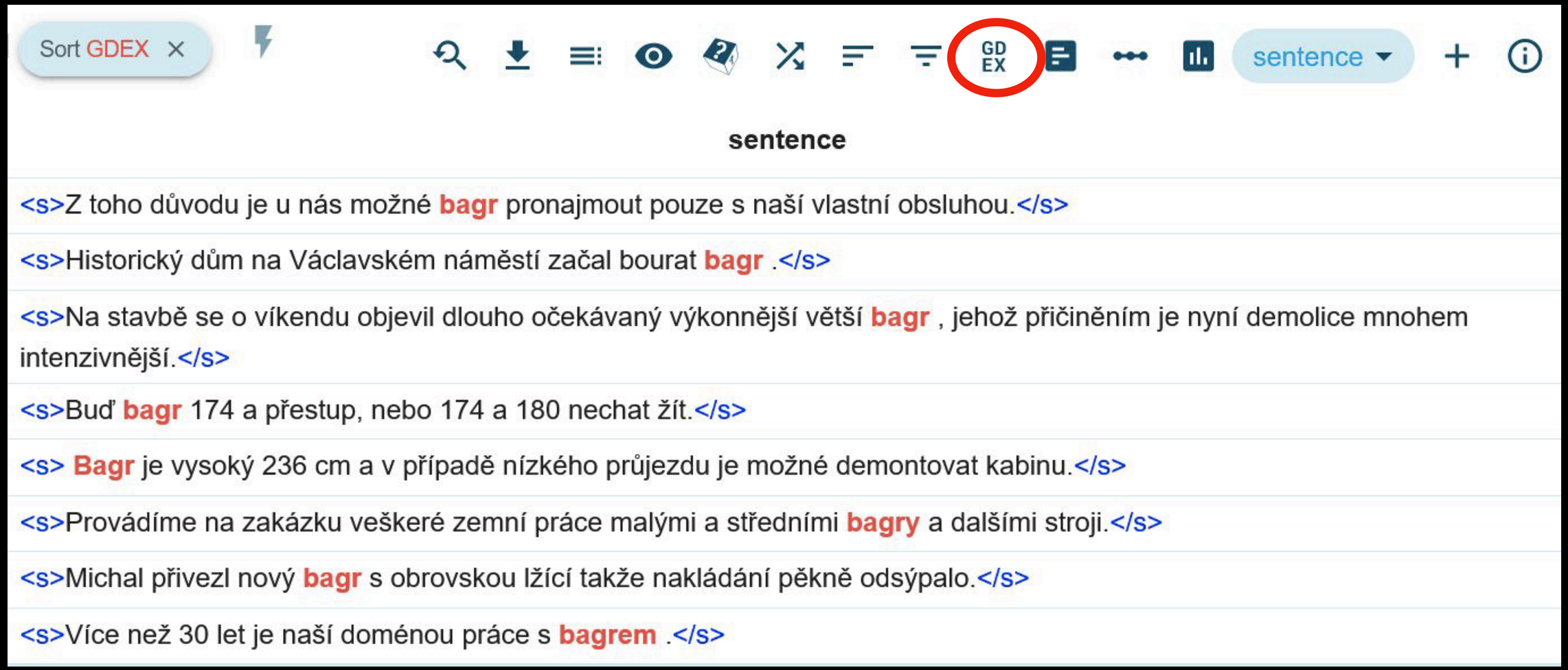
–> Jan si dělá starosti.

Nedobré příklady

„štamgast“

–> Jan je štamgast – chodí si k nám dát jedno či dvě skoro každou noc.

GDEX



sentence
<s>Z toho důvodu je u nás možné **bagr** pronajmout pouze s naší vlastní obsluhou.</s>
<s>Historický dům na Václavském náměstí začal bourat **bagr** .</s>
<s>Na stavbě se o víkendu objevil dlouho očekávaný výkonnější větší **bagr** , jehož přičiněním je nyní demolice mnohem intenzivnější.</s>
<s>Buď **bagr** 174 a přestup, nebo 174 a 180 nechat žít.</s>
<s> **Bagr** je vysoký 236 cm a v případě nízkého průjezdu je možné demontovat kabinu.</s>
<s>Provádíme na zakázku veškeré zemní práce malými a středními **bagry** a dalšími stroji.</s>
<s>Michal přivezl nový **bagr** s obrovskou lžící takže nakládání pěkně odsýpalo.</s>
<s>Více než 30 let je naší doménou práce s **bagrem** .</s>

Opáčko

Absolutní (korpusová?) frekvence

Relativní frekvence

Dokumentová frekvence

Opáčko – frekvence

Absolutní	ARF
Relativní	ALDF
Dokumentová	DP

Opáčko – frekvence

ARF (average reduced frequency)

ALDF (average logarithm distance frequency)

DP (deviation of proportions)

ALDF (average logarithm distance frequency)

Distribuce slova přes celý korpus.

ALDF blízko absolutní

=> obecně distribuované slovo.

ALD (average logarithm distance)

Vzdálenost mezi výskyty.

Přehled

Dobrý příklad dobře poznáte.

Editovat, či needitovat?

ARF, ALDF & další.