PLIN021 SÉMANTICKÁ ANALÝZA V PRAXI

ZUZANA NEVĚŘILOVÁ

2020/21

LEXIKÁLNÍ DESAMBIGUACE

SHRNUTÍ DOSAVADNÍCH ZNALOSTÍ

- Lexikální desambiguace je přirazení slova v daném kontextu významu v určitém repozitáři významů (např. Slovníku).
- **Leskův algoritmus** řeší úlohu porovnáním kontextu slova se slovy z definice významu.
- Úspěch algorimu silně závisí na použitém slovníku. Slovníky ale nebyly napsány s cílem být zdrojem pro algoritmus WSD.
- **Yarowského algoritmus** pracuje s korpusovými daty: **keyword in context (KWIC)**
- Trénovací data se musí pořídit ručně nebo ze slovníku kolokací. Stačí jich ale menší množství.
- Yarowského algoritmus nemusí rozhodnout všechny výskyty slova v kontextu.
- To ale lidé také pokaždé nedokážou.
- Je WSD vůbec **dobře definovaná úloha**?

LEXIKÁLNÍ DESAMBIGUACE: INVENTÁŘ VÝZNAMŮ

- Implicitní informace
- Granularita
- Související významy
- Aktuálnost, stáří

Výkladový slovník

- Co udělat jinak oproti výkladovému slovníku?

Specializovaný inventář

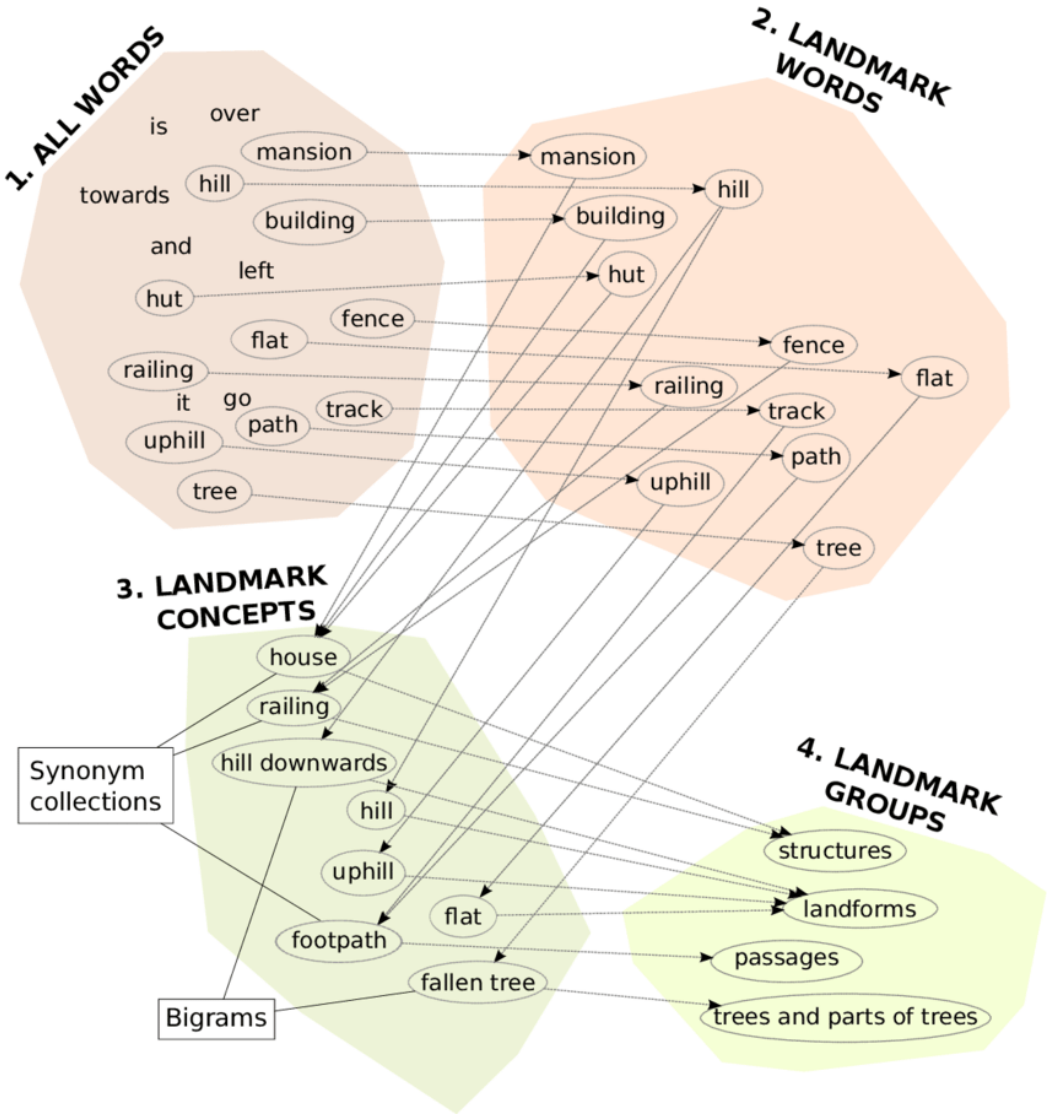
- Nestrukturovaná data
- Méně kontroly
- Jsou v korpusu všechna užití?

Korpus

- Jsou významy diskrétní?
- Potřebujeme významy rozlišit vždy?
- Nestačí nám někdy vědět, že slovo s v kontextu c má jiný vyznam, tj. $s_A(w, c_A) \neq s_B(w, c_B)$?

LEXIKÁLNÍ ZDROJE PRO NLP

- Historicky: slovníky
  - Nevýhody: chybějící explicitní informace
  - Vysoká cena → ne moc častá aktualizace
- Později až do současnosti: korpusy
  - Nevýhody: chyby v anotaci
  - Vyváženost? Reprezentativnost?
- Ontologie / sémantické sítě



https://www.researchgate.net/publication/277248905_Analysing_landmarks_in_nature_and_elements_of_geospatial_images_to_support_wayfinding/figures?lo=1

ONTOLOGIE – SÉMANTICKÉ SÍTĚ

Ontologie = explicitní specifikace sdílené konceptualizace (Gruber, 1995)

- Explicitní: co je řečeno, platí. Co není řečeno, neplatí.
- Sdílené: lidé by se na tom shodli.
- Konceptualizace: entity existují a existují i relace mezi nimi

- Implicitní informace
- Granularita
- Související významy
- Aktuálnost, stáří

Výkladový slovník

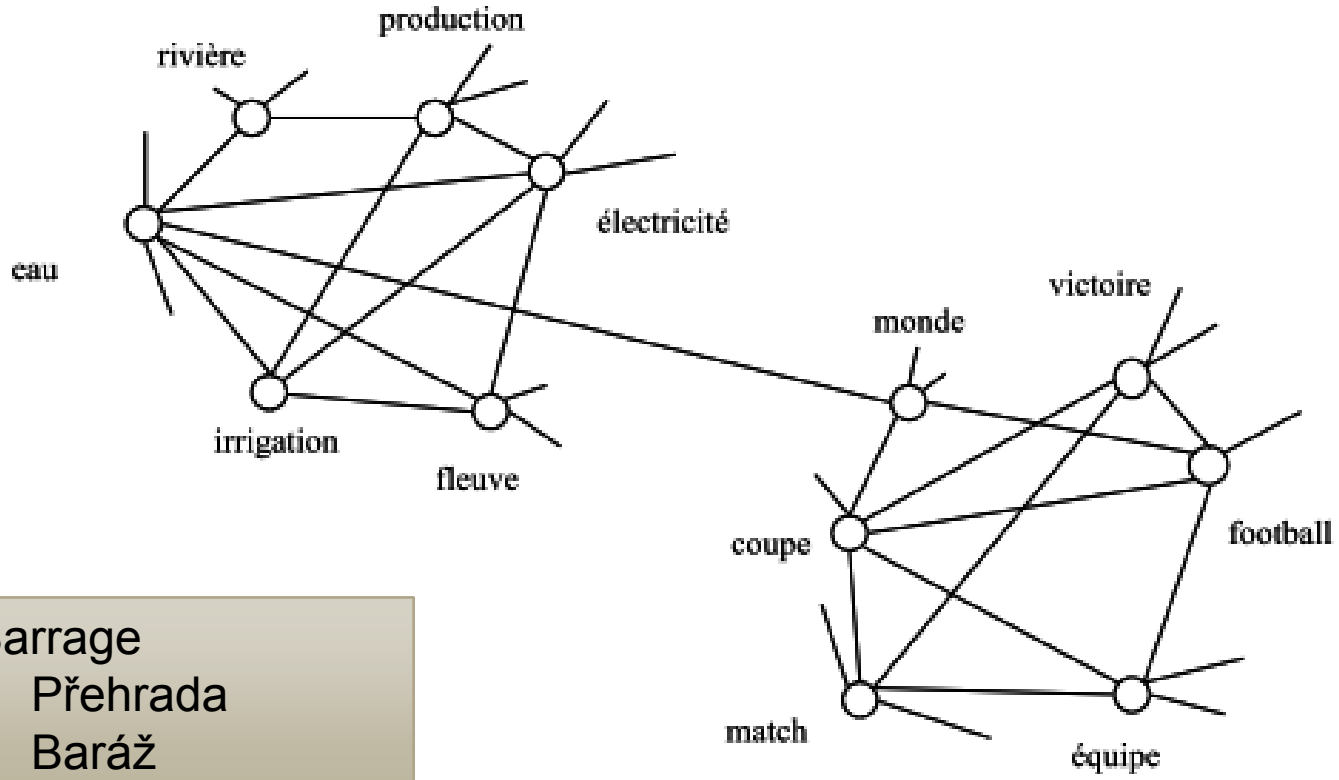
- Co udělat jinak oproti výkladovému slovníku?

Specializovaný inventář

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004

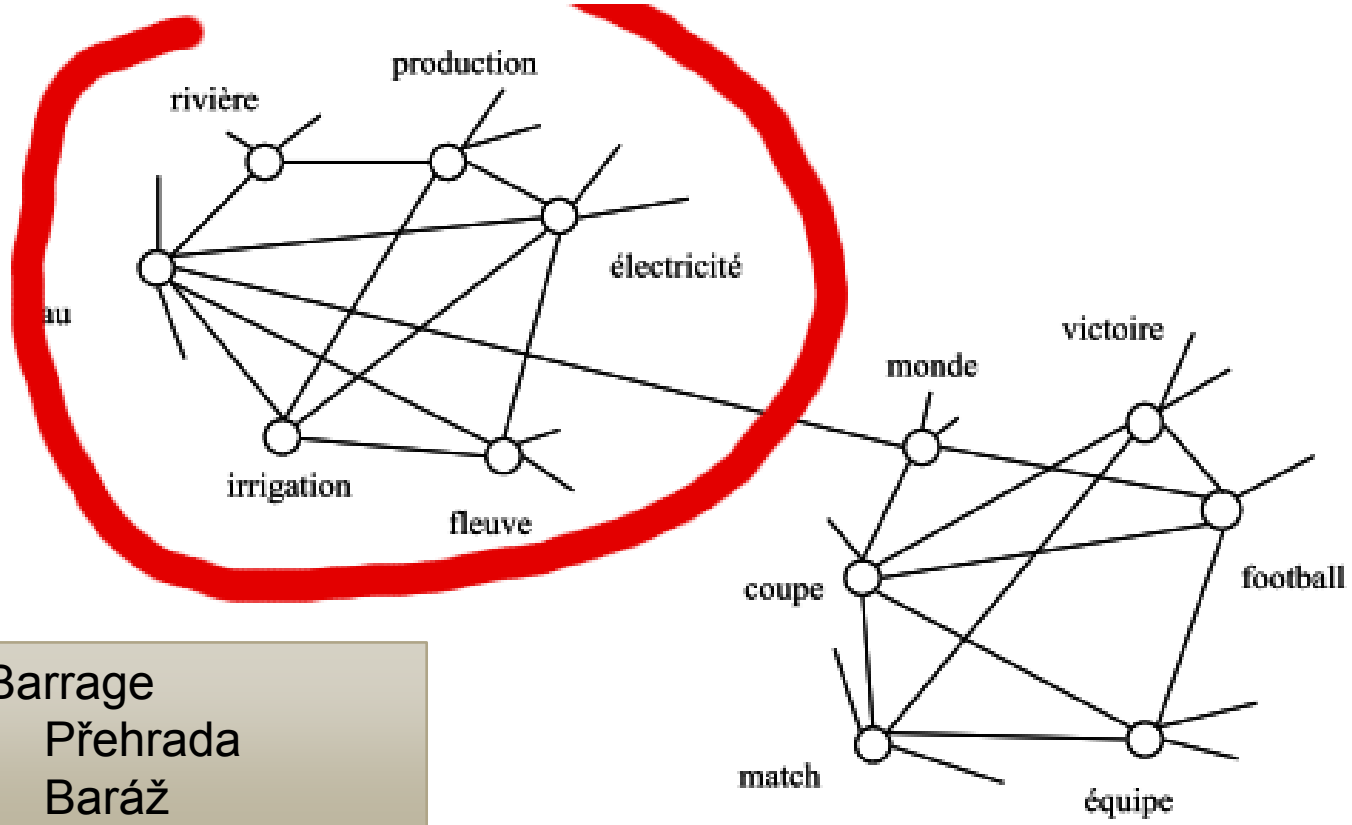
- Word sense disambiguation → word sense discrimination
- „malé světy“ (Milgram, 1967), nalezení centrálních uzlů, „influencerů“
- Graf sousednostI, vážené hrany A↔B mezI slovy A a B:
  - $w_{AB} = 0$, pokud se slova vyskytují vždy spolu
  - $w_{AB} = 1$, pokud se nikdy spolu nevyskytují
  - $w_{AB} = 1 - \max[p(A|B), p(B|A)]$
  - Hrany s $w_{AB} > 0{,}9$ se zahodí
- rozdělení grafu na podgrafy (NP-těžký problém)
  - Pro tisíce uzlů je jedinou rozumnou cestou aproximativní, heuristický algoritmus.

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004



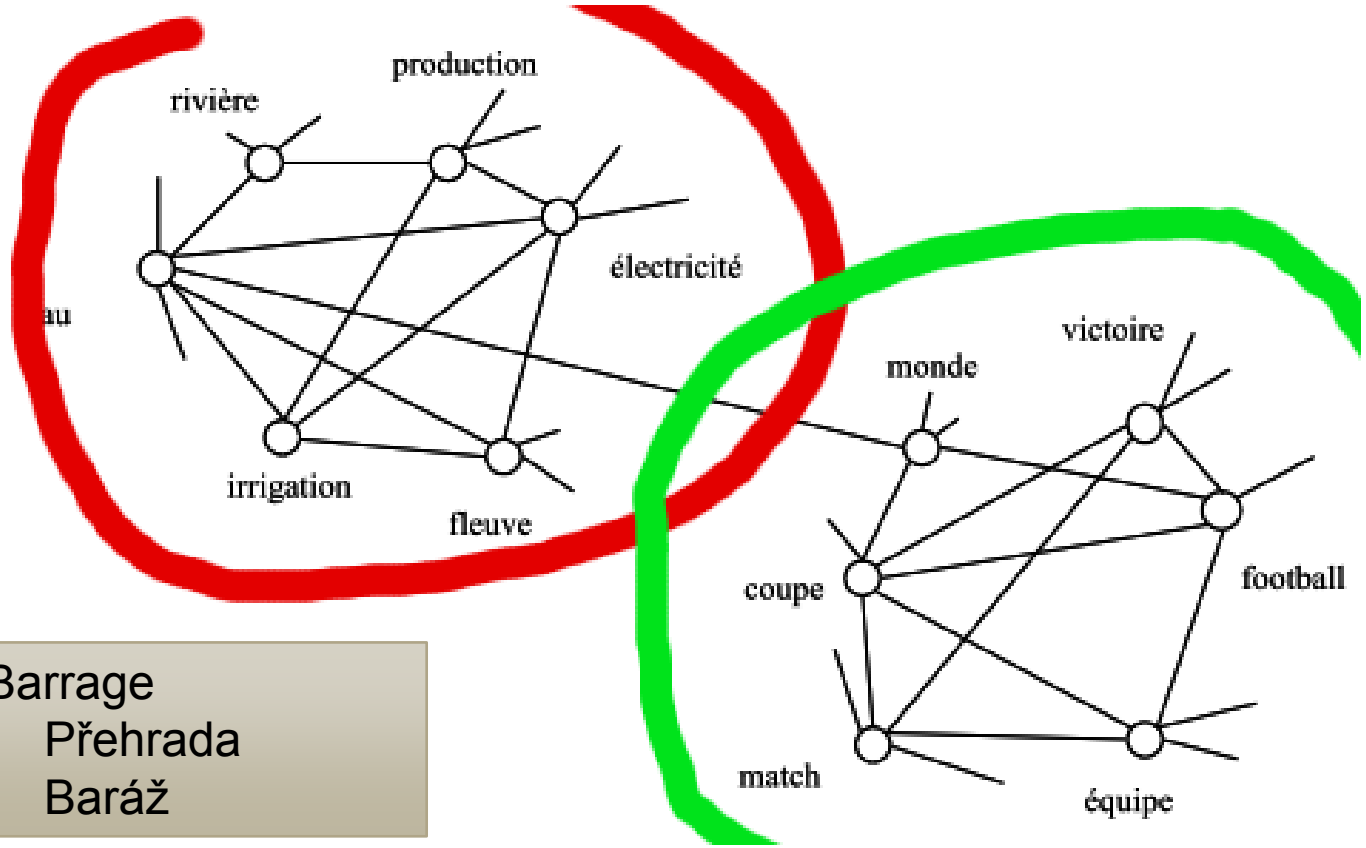
Barrage
- Přehrada
- Baráž

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004



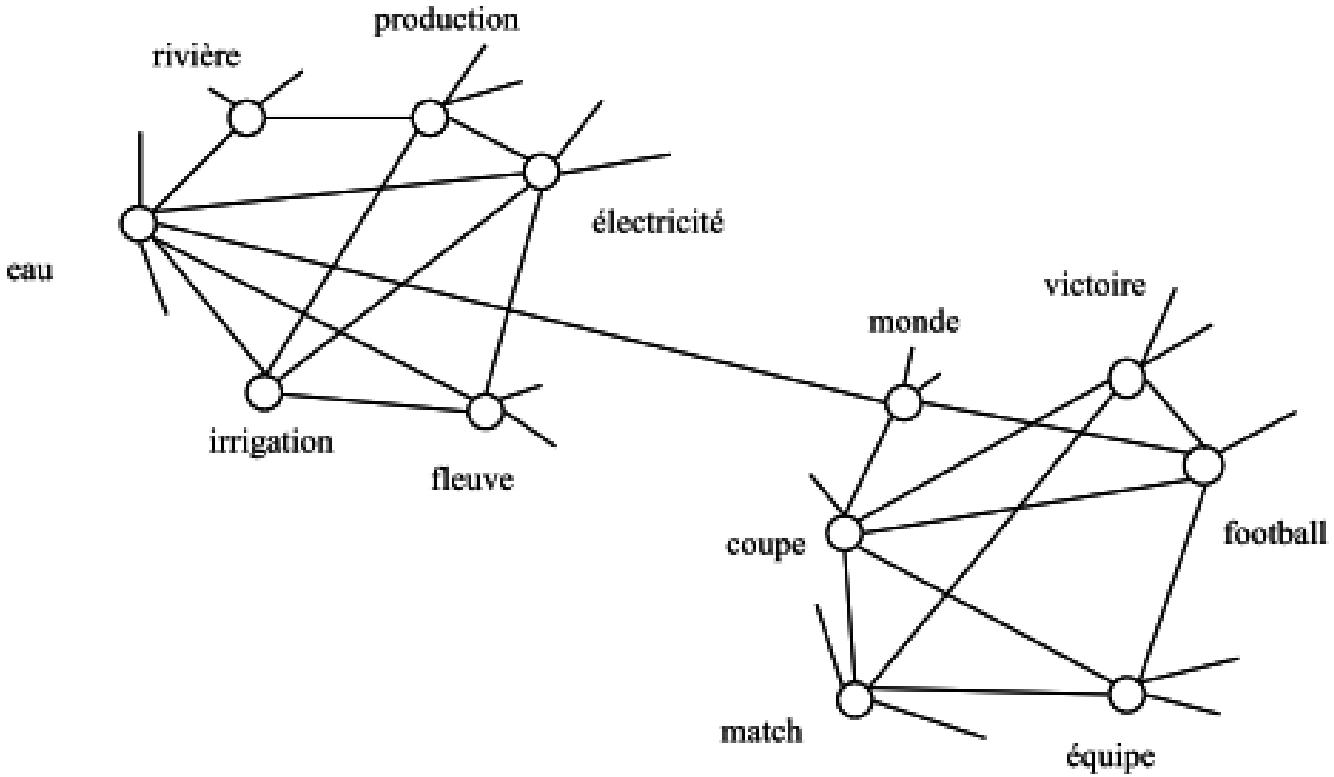
Barrage
- Přehrada
- Baráž

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004

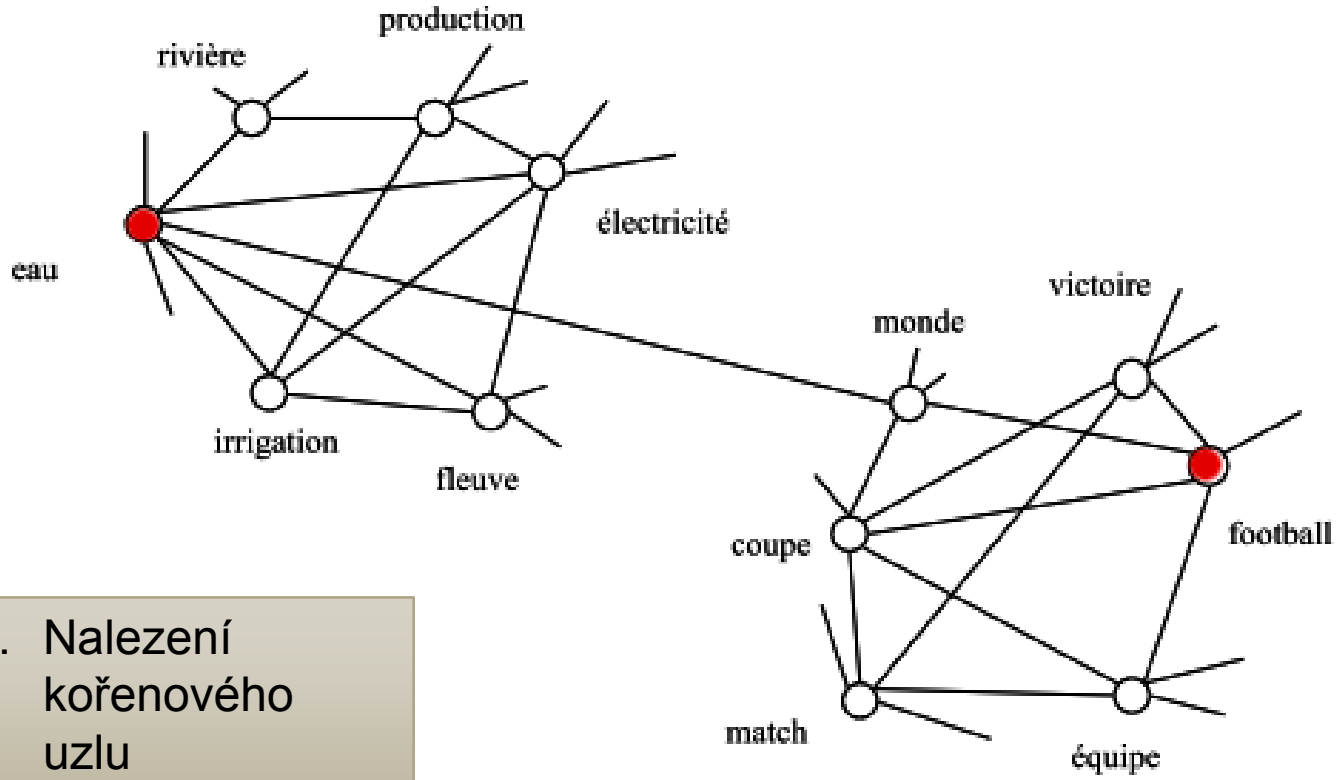


Barrage
- Přehrada
- Baráž

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004

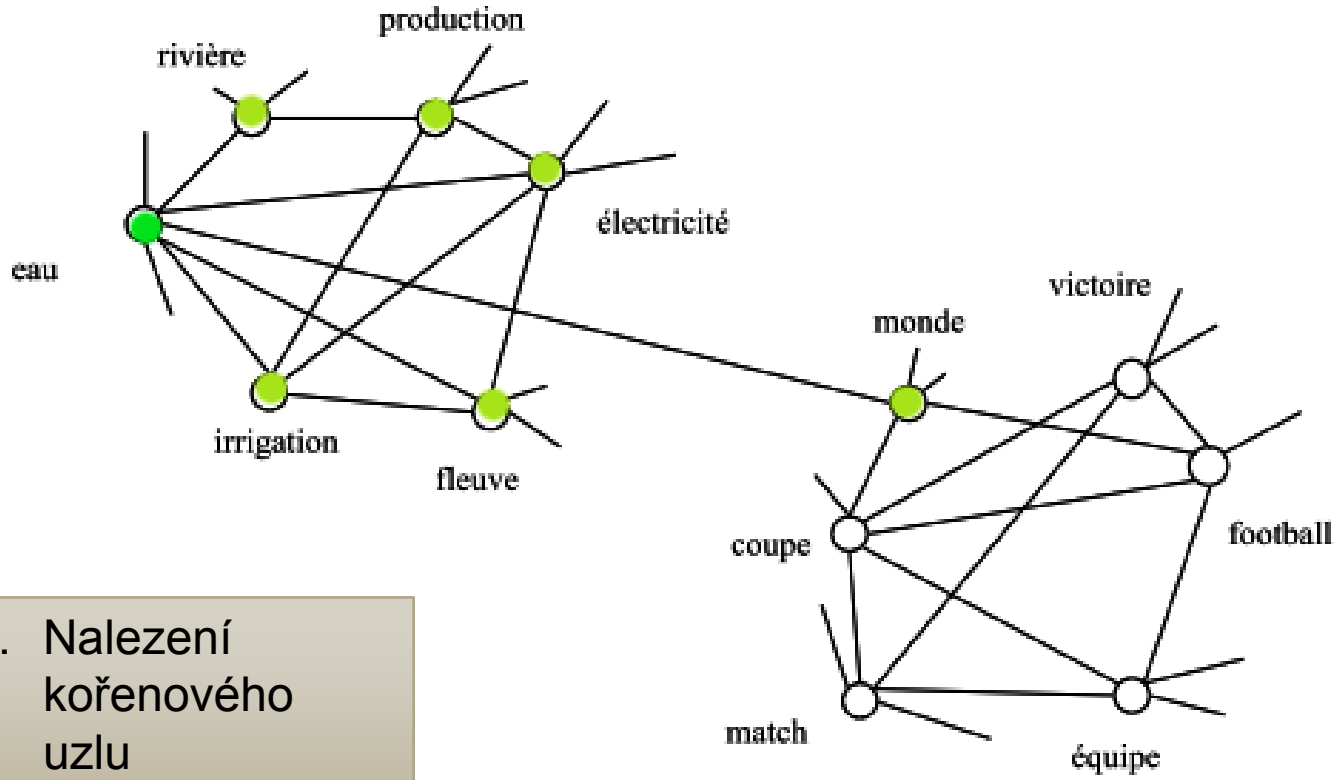


WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004



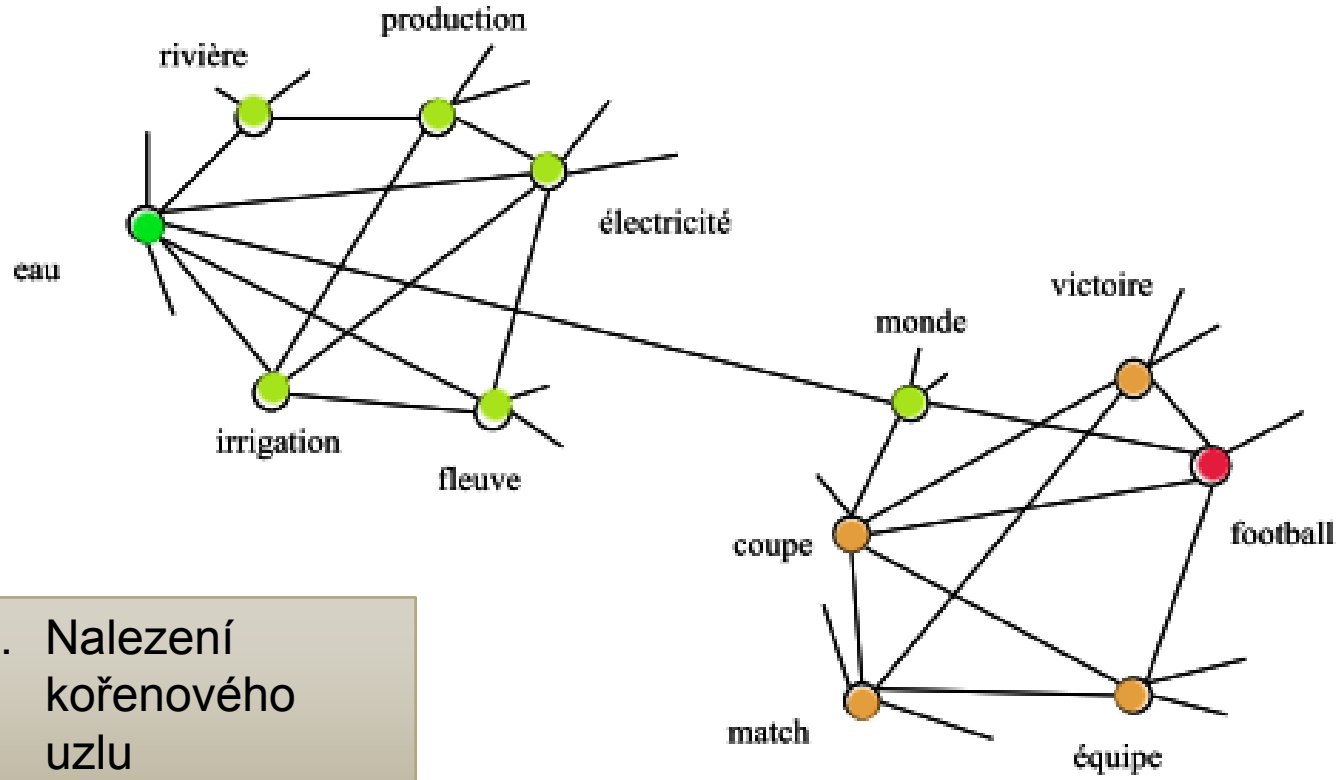
1. Nalezení kořenového uzlu
2. Nalezení minimální kostry

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004



1. Nalezení kořenového uzlu
2. Nalezení minimální kostry

WSD BEZ INVENTÁŘE VÝZNAMŮ – HYPERLEX, 2004



1. Nalezení kořenového uzlu
2. Nalezení minimální kostry

WORD SENSE DISCRIMINATION

WORD SENSE INDUCTION

- Grafy spoluvýskytu (sousednost, spoluvýskyt v pevně daném okně kontextu)
  Co-occurrence graphs
- Slovní klastry
  Word clusters
- Kontextové klastry
  Context clusters

WSD – ÚSPĚCH? KVALITA?

Soutěž SENSEVAL (www.senseval.org)

- vyhodnocení systémů pro WSD
- od roku 1998 (Senseval-1, -2, -3, Semeval-2007, -2010)
- od Semeval-1 jsou úkoly různé (např analýza sentimentu, detekce metonymie)
- data z proběhlých kol jsou k dispozici

Soutěž SemEval

- 2010 WSI (word sense induction)
- 2013 Task 12, 2015 Task 13 - Multilingual WSD

LITERATURA

- Gruber, Thomas. **Toward Principles for the Design of Ontologies Used for Knowledge Sharing**. *International Journal Human-Computer Studies* Vol. 43, Issues 5-6, Novemer 1995, p.907-928. https://tomgruber.org/writing/onto-design.pdf
- Ide N., Wilks Y. (2007) **Making Sense About Sense**. In: Agirre E., Edmonds P. (eds) Word Sense Disambiguation. Text, Speech and Language Technology, vol 33. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4809-8_3. http://staffwww.dcs.shef.ac.uk/people/Y.Wilks/papers/ide-wilks_final.pdf
- Jean Véronis, **HyperLex: lexical cartography for information retrieval,** Computer Speech & Language, Volume 18, Issue 3, 2004, pages 223-252, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2004.05.002. http://www.sciencedirect.com/science/article/pii/S0885230804000142