PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

31. října 2017

Word Sense Disambiguation/Discrimination

Reprezentace znalostí

Sémantické role

Sémantické role a slovesa

Sémantické rámce

Sémantické sítě

Word Sense Disambiguation

pro dané slovo: nalezení čísla významu (z inventáře významů) na základě kontextu

Word Sense Disambiguation: slabiny

největší slabinou je inventář významů

Word Sense Disambiguation: slabiny

největší slabinou je inventář významů

proto existují snahy úplně se inventářům vyhnout

Word Sense Discrimination

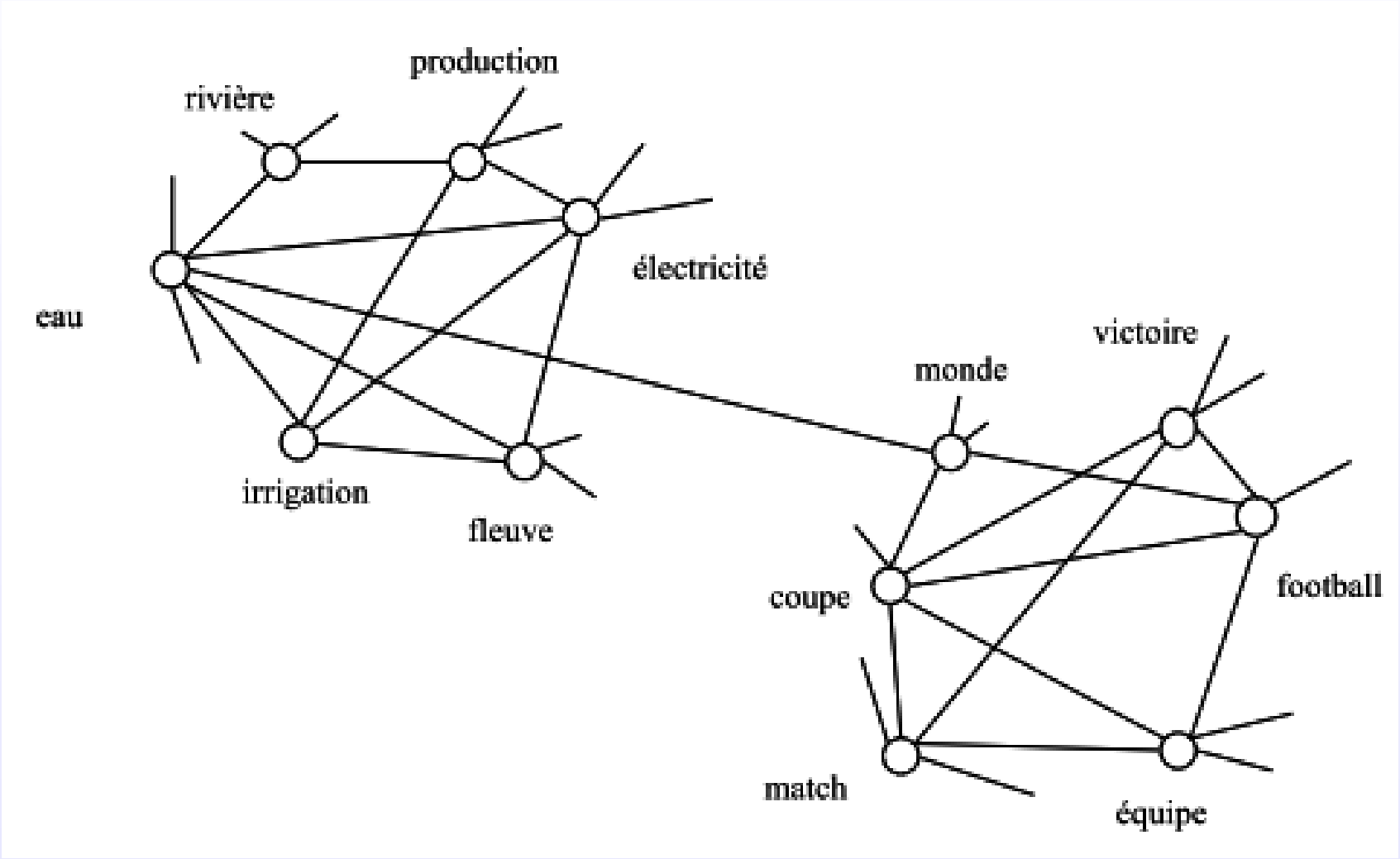
pro dané slovo: rozdělení významů na klastry (bez inventáře významů), pravděpodobnost příslušnosti k určitému klastru na základě kontextu

- rozhodovací strom
- naivní Bayesovský klasifikátor
- jiný algoritmus

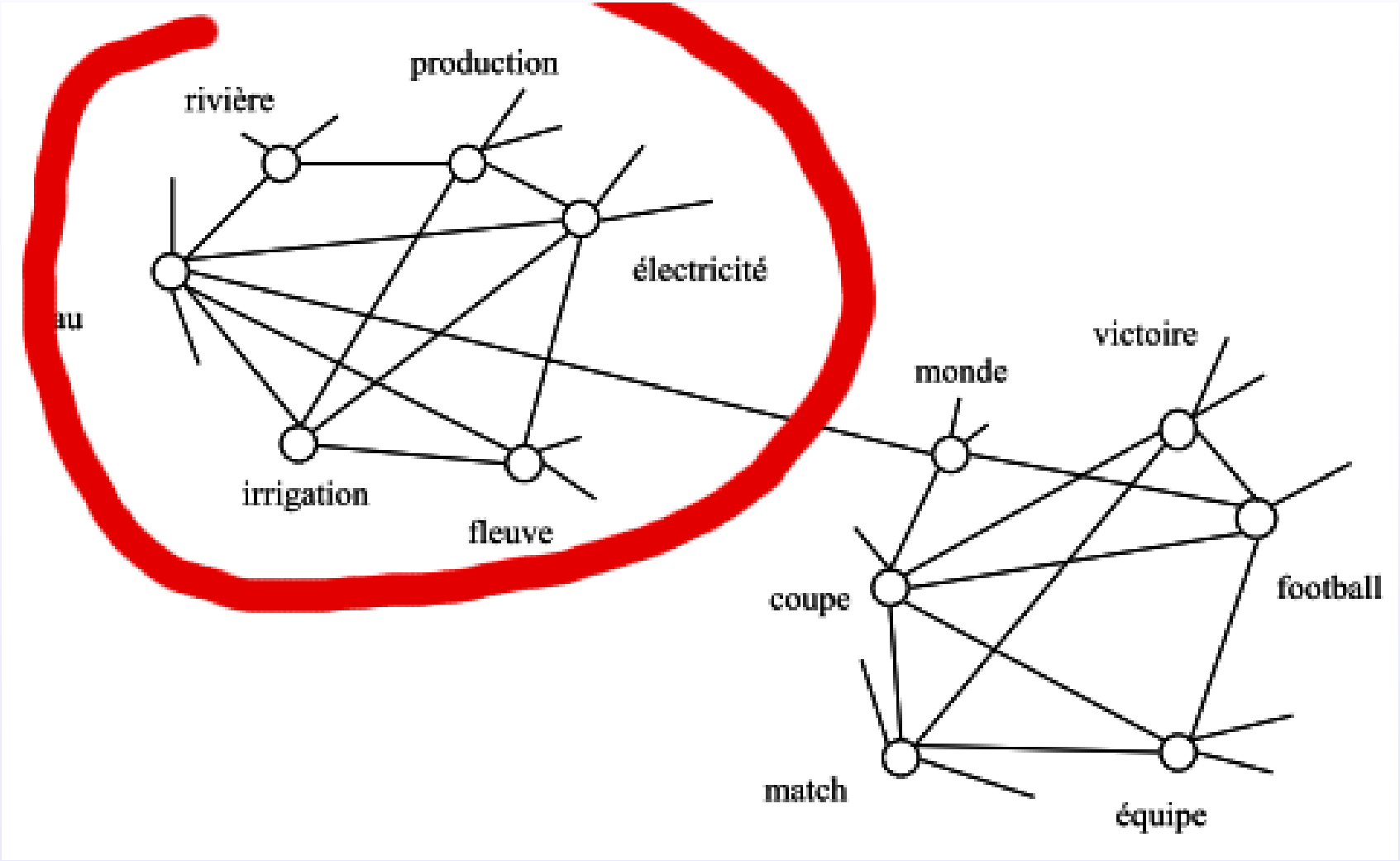
Word Sense Discrimination: grafy

- „malé světy“ (Milgram, 1967)
- graf (HyperLex, [Véronis, 2004])
- vážené hrany A–B:
  - $w = 0$, pokud se slova vyskytují vždy spolu
  - $w = 1$, pokud se nikdy spolu nevyskytují
  - $w_{AB} = 1 - max[p(A|B), p(B|A)]$
- rozdělení grafu na podgrafy (NP-těžký problém)

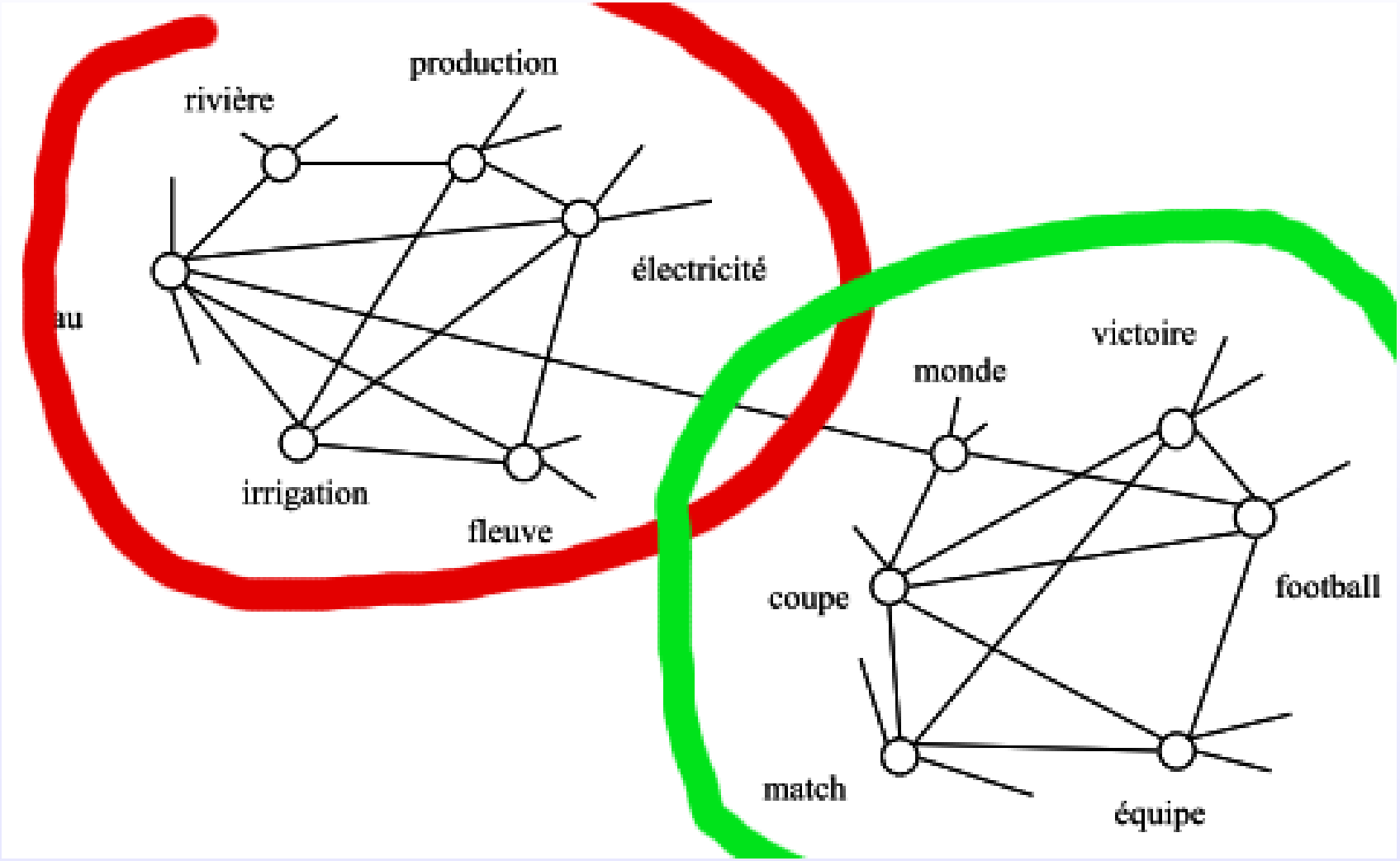
HyperLex: příklad „barrage“



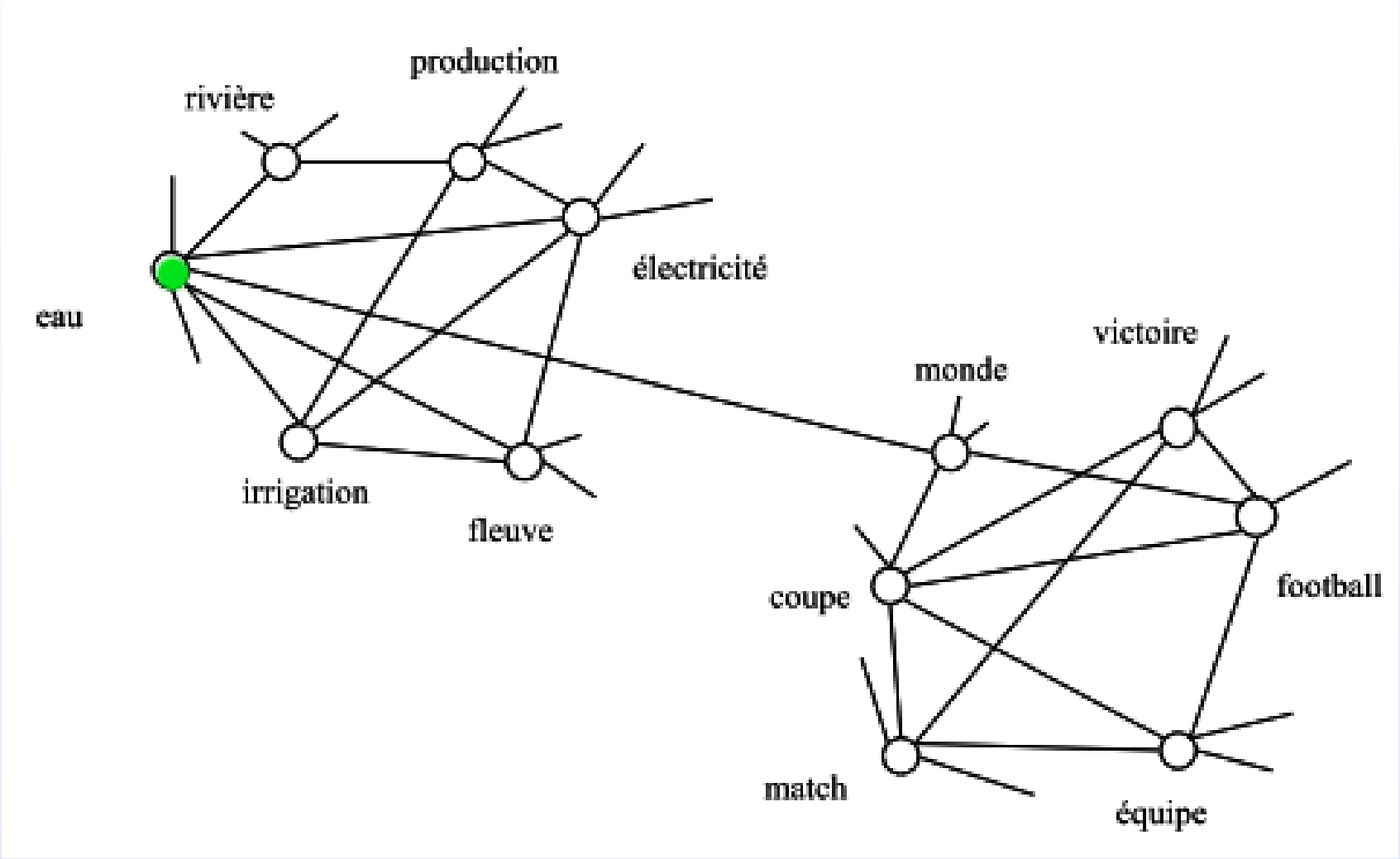
HyperLex: příklad „barrage“



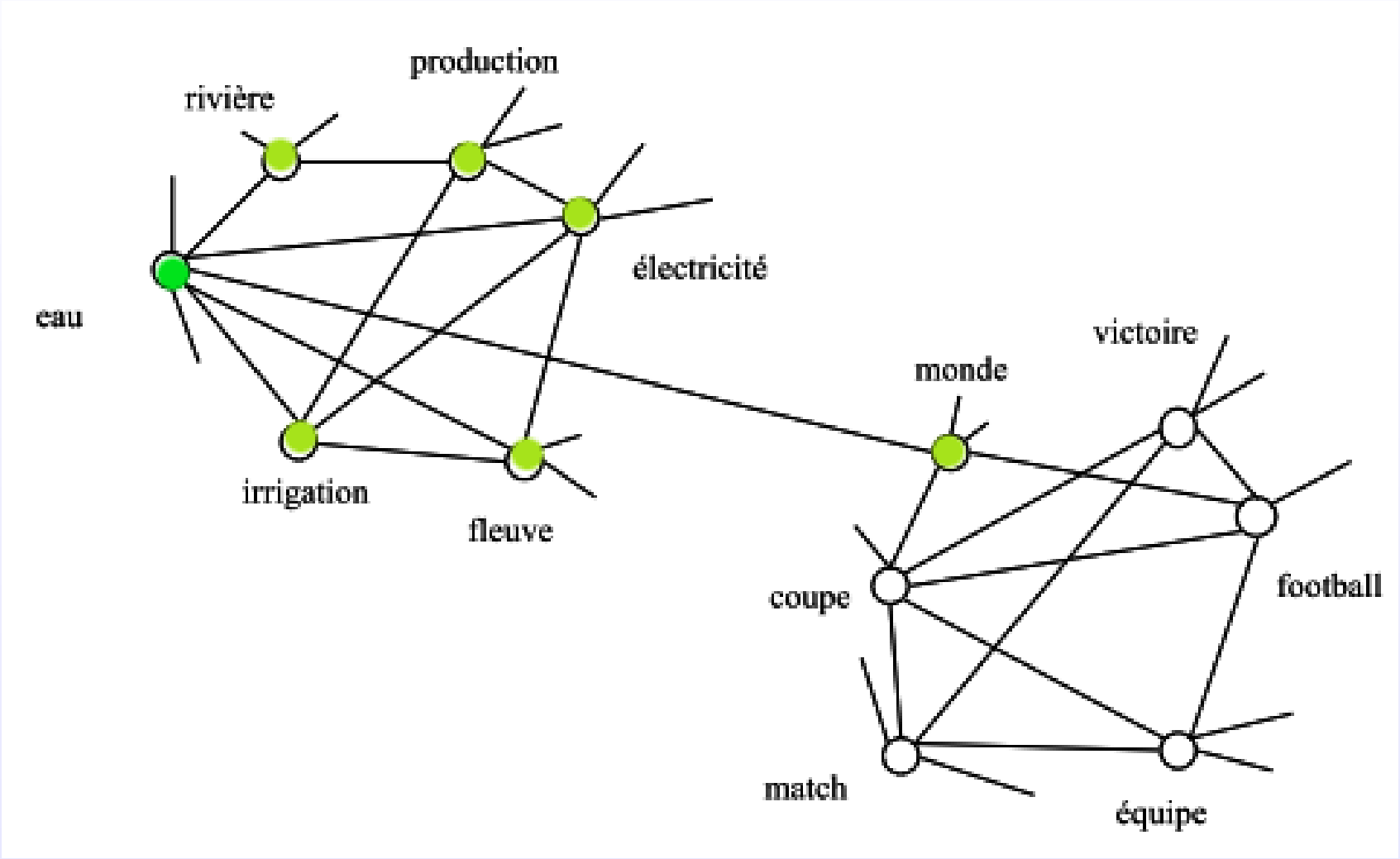
HyperLex: příklad „barrage“



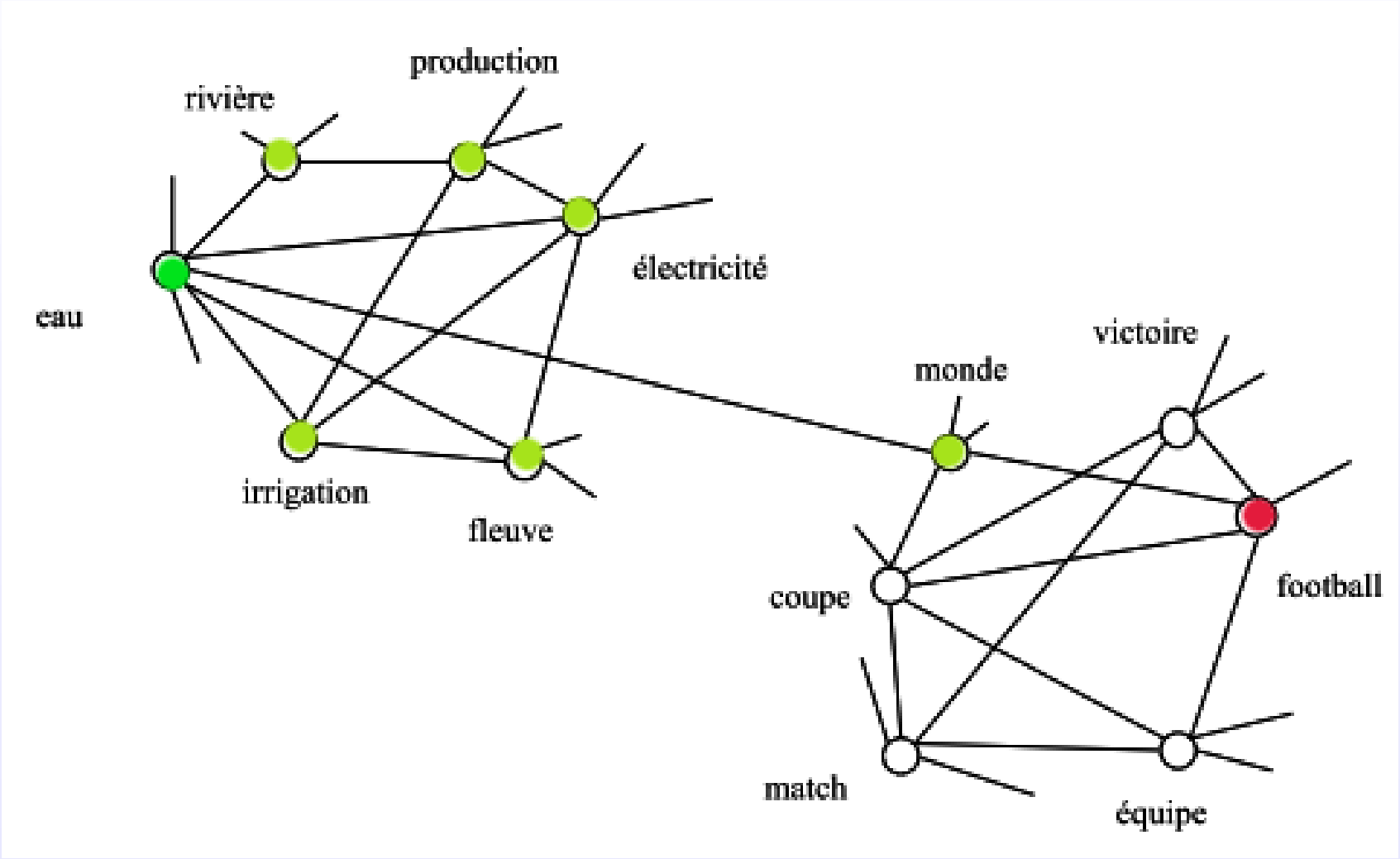
HyperLex: nalezení kořenového uzlu



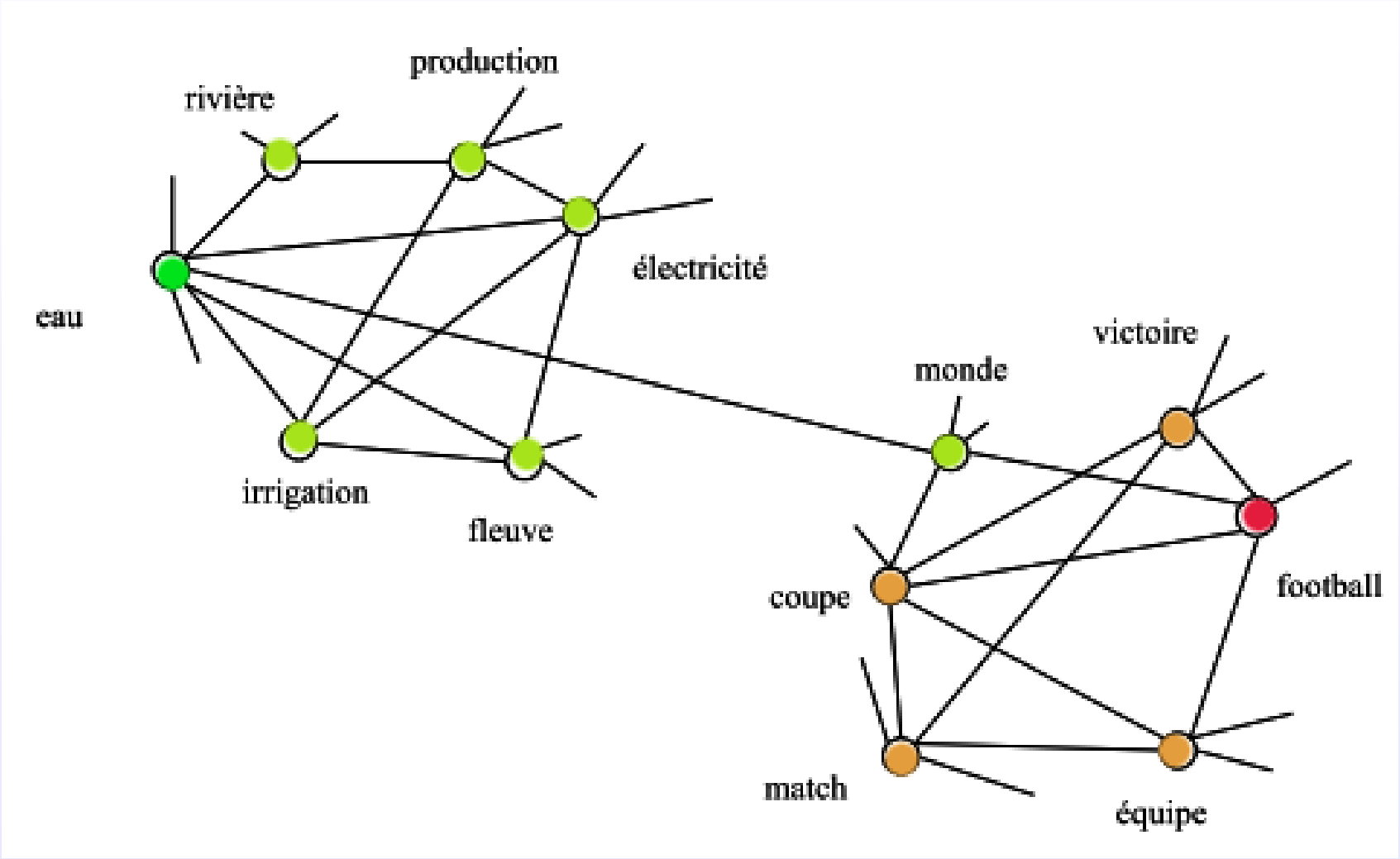
HyperLex: nalezení kořenového uzlu



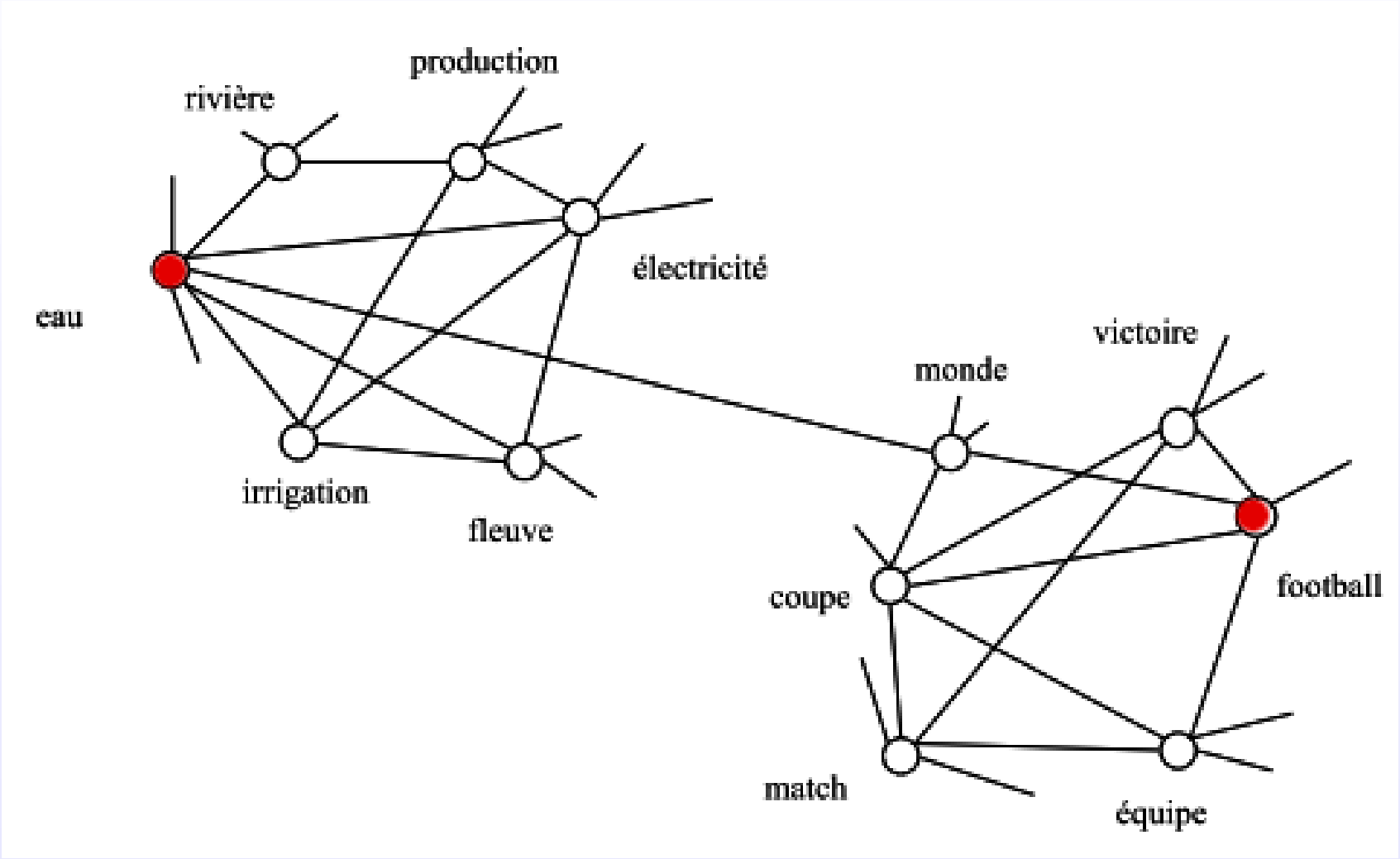
HyperLex: nalezení kořenového uzlu



HyperLex: nalezení kořenového uzlu



HyperLex: nalezení minimální kostry



Grafové klastrování

<https://nlp.fi.muni.cz/projekty/visualbrowser/wsd/>

Reprezentace znalostí

- sémantické rysy:
  matka = FEMALE + ADULT + HAS CHILD
  batole = HUMAN − ADULT
- výběrová omezení:
  Rodinné domy postaví malé stavební firmy.
- sémantické role
- sémantické rámce: VerbNet, VALLEX, VerbaLex, FrameNet
- sémantické sítě: WordNet, ConceptNet

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

- Koupila jsem si pletenou čepici a šálu.
- Koupila jsem si nealkoholické pivo a křupky.

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

- Koupila jsem si pletenou čepici a šálu.
- Koupila jsem si nealkoholické pivo a křupky.

pletená šála – OK, nealkoholické křupky – NOT OK

Výběrová omezení (selectional restrictions)

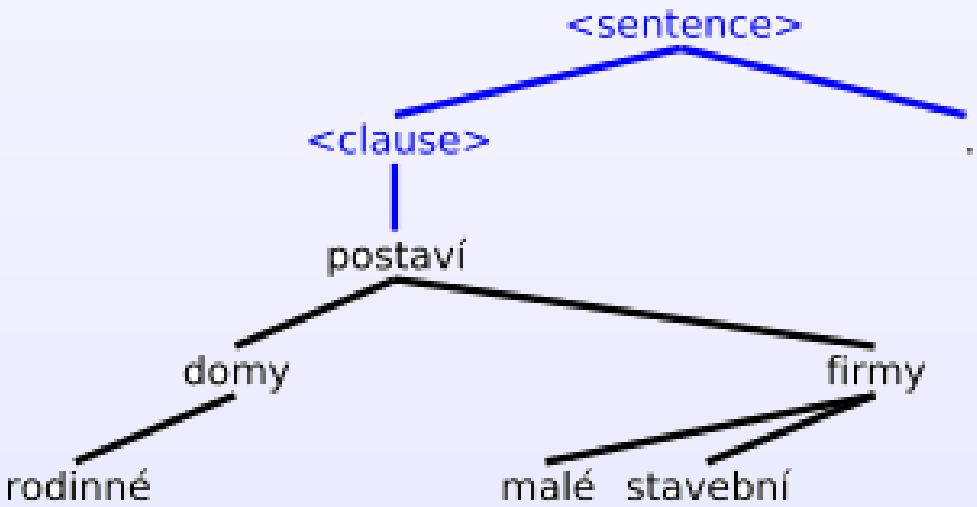
slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

Rodinné domy postaví malé stavební firmy.

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

Rodinné domy postaví malé stavební firmy.



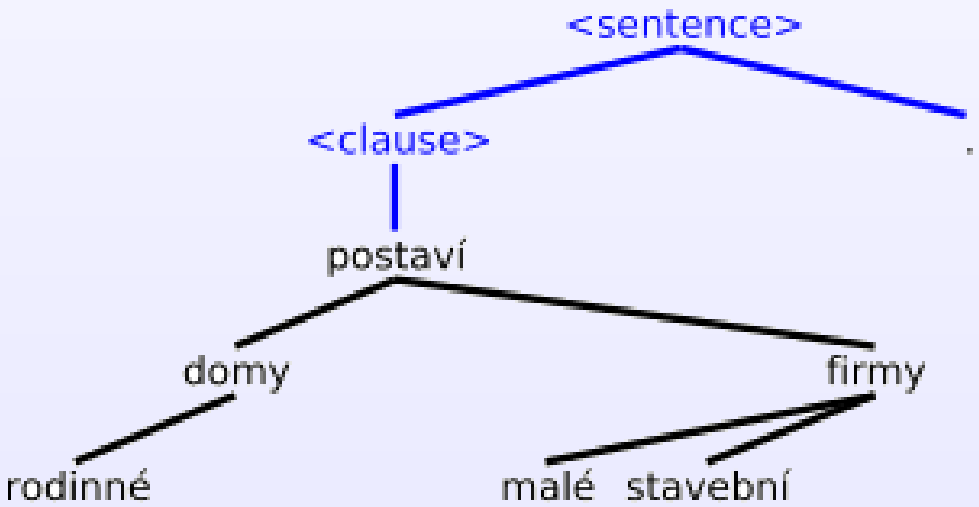
AGENT = rodinné domy
THEME = malé stavební firmy

AGENT = malé stavební firmy
THEME = rodinné domy

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

Rodinné domy postaví malé stavební firmy.



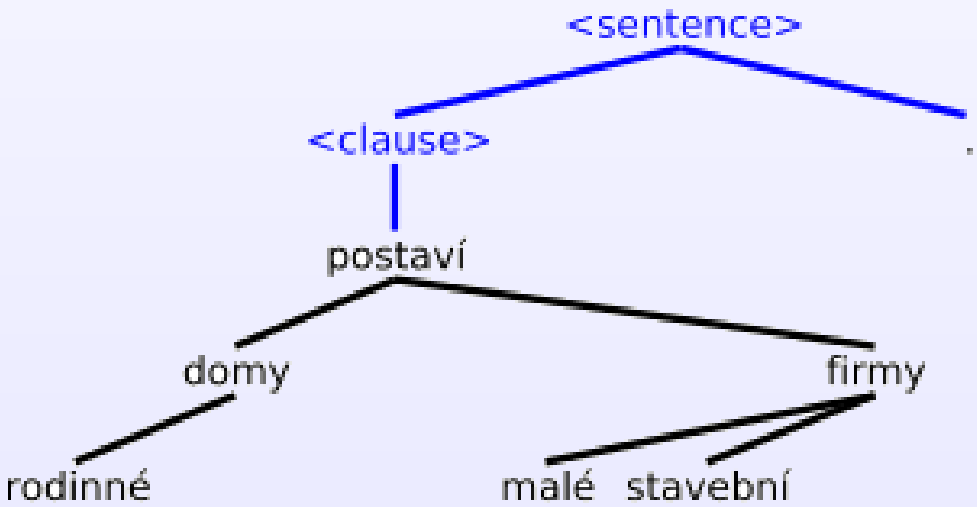
AGENT = rodinné domy
THEME = malé stavební firmy

AGENT = malé stavební firmy
THEME = rodinné domy

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

Rodinné domy postaví malé stavební firmy.



AGENT = rodinné domy
THEME = malé stavební firmy

AGENT = malé stavební firmy
THEME = rodinné domy

(AGENT postavit PERSON | INSTITUTION)
(THEME postavit BUILDING)

Sémantické role (semantic role, thematic relation, theta-role, deep case)

(AGENT postavit PERSON | INSTITUTION)
(THEME postavit BUILDING)

Sémantické role (semantic role, thematic relation, theta-role, deep case)

(AGENT postavit PERSON | INSTITUTION)
(THEME postavit BUILDING)

AGENT, EXPERIENCER, THEME, PATIENT, INSTRUMENT, FORCE/NATURAL CAUSE, LOCATION, DIRECTION/GOAL, RECIPIENT, SOURCE/ORIGIN, TIME, BENEFICIARY, MANNER, PURPOSE, CAUSE

Sémantické role (semantic role, thematic relation, theta-role, deep case)

(AGENT postavit PERSON | INSTITUTION)
(THEME postavit BUILDING)

AGENT, EXPERIENCER, THEME, PATIENT, INSTRUMENT, FORCE/NATURAL CAUSE, LOCATION, DIRECTION/GOAL, RECIPIENT, SOURCE/ORIGIN, TIME, BENEFICIARY, MANNER, PURPOSE, CAUSE

Jména rolí jsou u různých autorů mírně různá (např. AGENT/ACTOR). Přiřazení rolí je někdy zřejmé (Karel/AGENT rozbil okno), jindy zřejmé není (Kladivo/AGENT?INSTRUMENT? rozbilo okno).

[Fillmore, 1968]

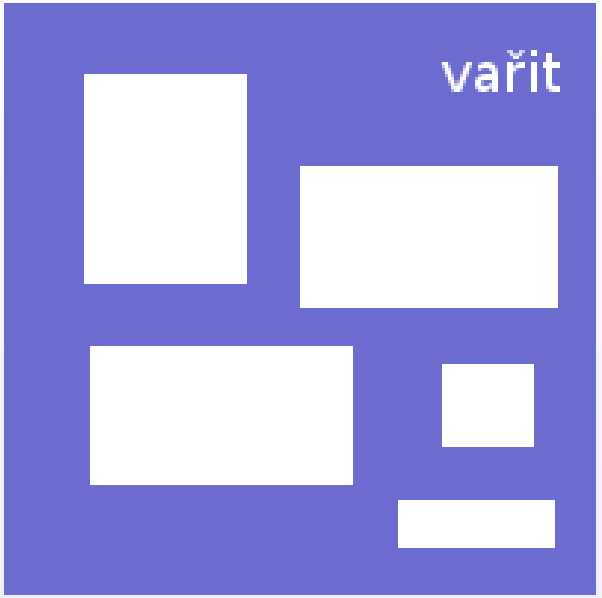
Sémantické role (semantic role, thematic relation, theta-role, deep case)

sémantické role [Jackendoff, 1992, kap. 2.2]

- jsou nositelé širšího významu
- jsou součástí sémantické struktury promluvy (conceptual structure), ne syntaxe
- typicky však odpovídají větným členům
- objevují se v sémantických rámcích (o těch později)

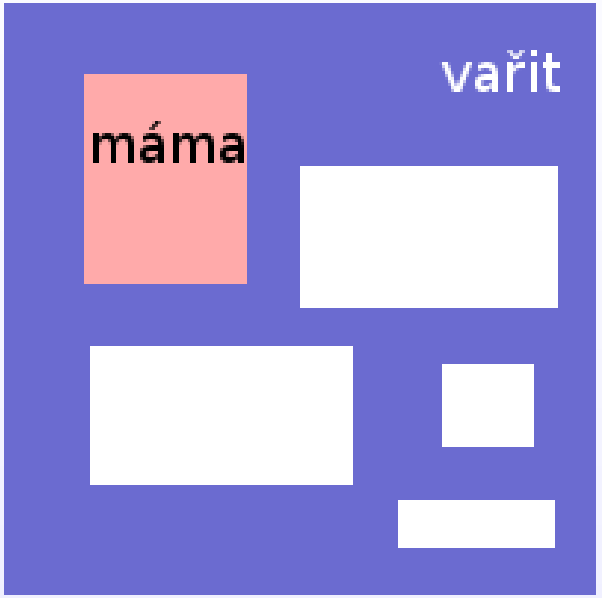
Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



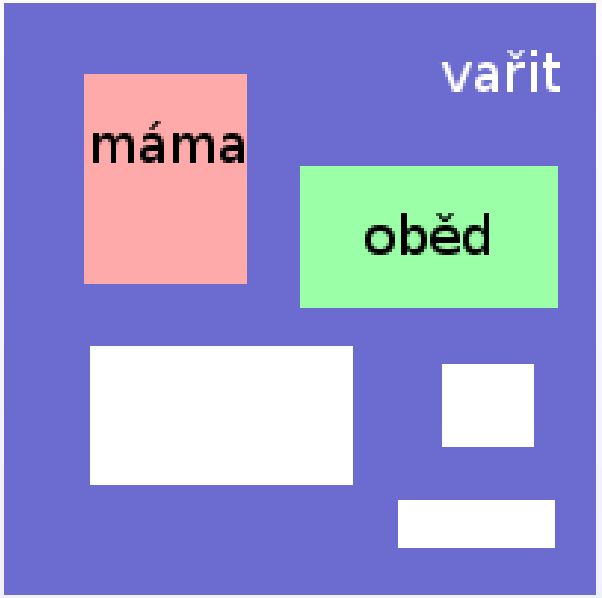
Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



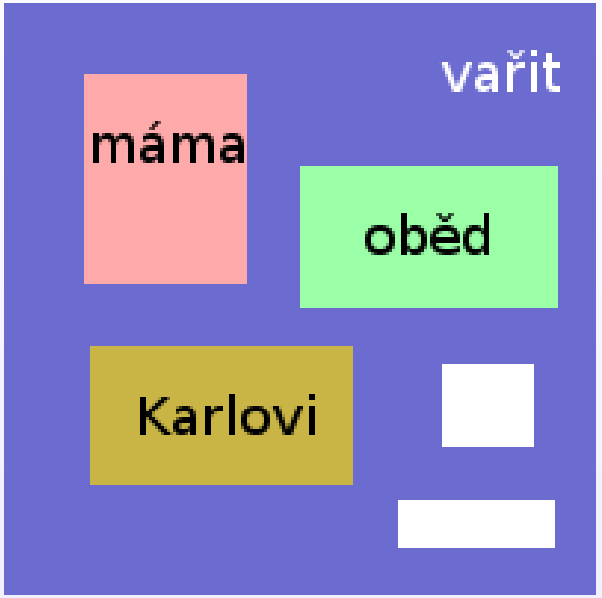
Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



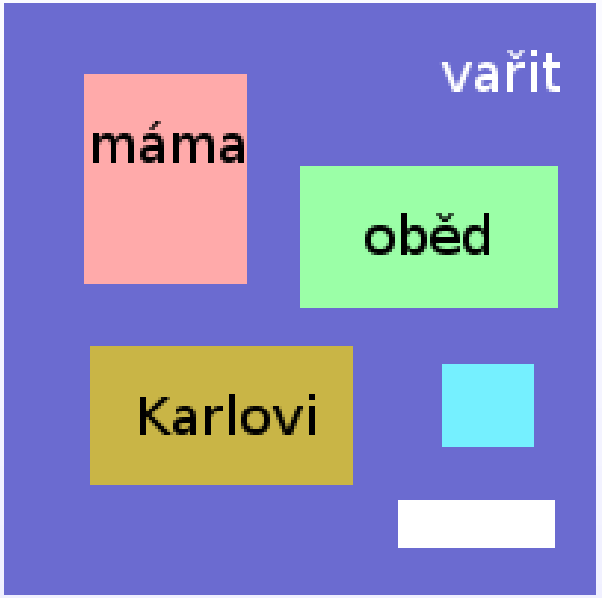
Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



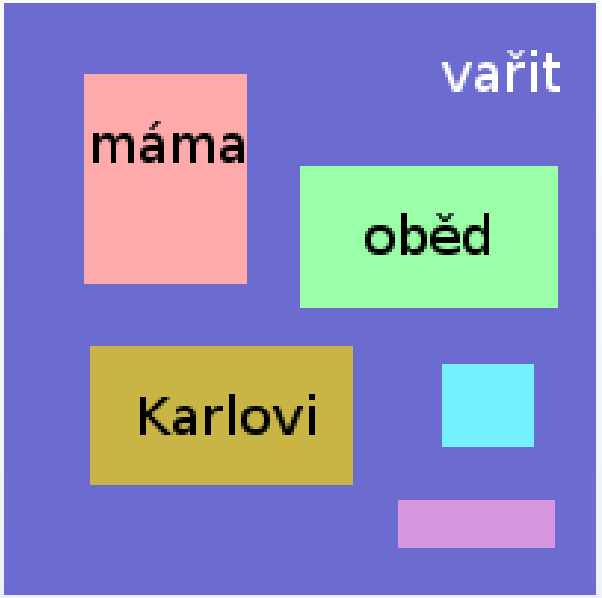
Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



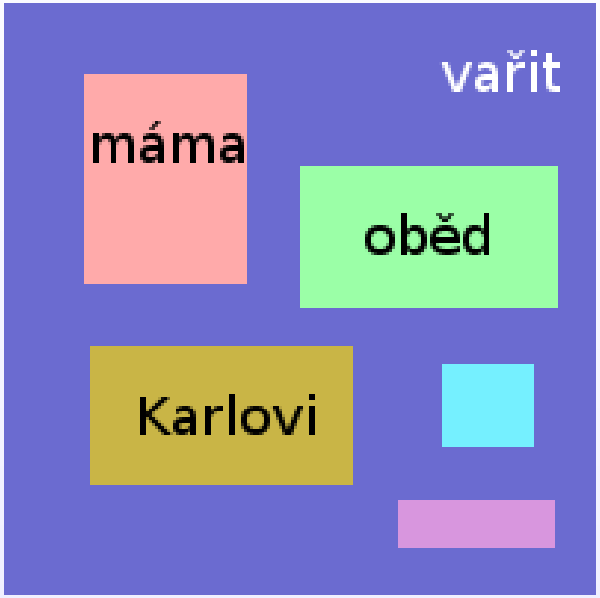
Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



Sémantické role ve slovesných valencích

sloveso je centrem věty – na slovese „visí“ celý zbylý význam
sloveso si lze představit jako rámec, do kterého zapadnou ostatní dílky



slovník slovesných valencí VerbaLex [Hlaváčková, 2007]

Rámce – stereotypická informace [Minsky, 1974]

objekty, vlastnosti, vztahy mezi objekty, odvozovací pravidla

pro každý objekt jsou v rámci rubriky *sloty (slots)*, každá rubrika má položky (*links, facets*) jako např. aktuální hodnotu, implicitní hodnotu, rozsah možných hodnot ...

Rámce – stereotypická informace [Minsky, 1974]

objekty, vlastnosti, vztahy mezi objekty, odvozovací pravidla

pro každý objekt jsou v rámci rubriky *sloty (slots)*, každá rubrika má položky (*links, facets*) jako např. aktuální hodnotu, implicitní hodnotu, rozsah možných hodnot ...

kočka domácí:

- je_druhem zvíře
- má_nepřítele pes
- má_za_potravu myš
- nachází_se (blízko_domu, doma)
- velikost malé_zvíře

Rámce – skutečné projekty

FrameNet, ConceptNet

slovesa – VerbNet, VerbaLex

Rámce – použití

Rámce můžeme použít pro desambiguaci slov i celých vět [Laparra and Rigau, 2009].

[Bernard Lansky]$_{STUDENT}$ *studied* [the piano]$_{SUBJECT}$
[with Peter Wallfisch]$_{TEACHER}$.

Rámce – použití

Rámce můžeme použít pro doplnění implicitní (nezmiňované) znalosti.

Koupila jsem ojetou felicii. Byly to vyhozené peníze.

koupit:

- má_činitele člověk/instituce/skupina
- má_benefaktora člověk/instituce/skupina
- má_předmět výrobek/nemovitost/zvíře/rostlina/přírodnina
- má_část činitel dá peníze
- má_část benefaktor dá předmět

Rámce – použití

Rámce můžeme použít pro doplnění implicitní (nezmiňované) znalosti.

Koupila jsem ojetou felicii. Byly to vyhozené peníze.

koupit:

- má_činitele člověk/instituce/skupina
- má_benefaktora člověk/instituce/skupina
- má_předmět výrobek/nemovitost/zvíře/rostlina/přírodnina
- má_část činitel dá peníze
- má_část benefaktor dá předmět

Rámce – použití

Rámce můžeme použít pro doplnění implicitní (nezmiňované) znalosti.

Koupila jsem ojetou felicii. Byly to vyhozené peníze.

koupit:

- má_činitele člověk/instituce/skupina
- má_benefaktora člověk/instituce/skupina
- má_předmět výrobek/nemovitost/zvíře/rostlina/přírodnina
- má_část činitel dá peníze
- má_část benefaktor dá předmět

Sémantické sítě

- uzly: pojmy (mnoho uzlů)
- relace: vztahy mezi pojmy (málo druhů relací)
- předpoklady: jeden pojem v jednom uzlu, v jednom uzlu jeden pojem
- odvozování v sémantických sítích díky tranzitivitě některých relací (kterých?)

WordNet a EuroWordNet

WordNet (Princeton WordNet, PWN) – lexikální síť

- původně nástroj k ověření teorie o uspořádání lidské paměti (G. A. Miller, od r. 1985)
- počítačově dobře zpracovatelný zdroj informací o významech slov a vztazích mezi významy [Fellbaum, 1998]
- jednotkou je synonymická řada (synonymical set, synset)
- synsety jsou spojeny relacemi:
  - hyperonymie/hyponymie: vůz, automobil – dodávka
  - holonymie/meronymie (part of, member of): vůz, automobil – tlumič; orchestr – houslista
  - troponymie: šeptat – mluvit
  - near-antonym: den – noc
  - odvození: velikost – velký
- slovní druhy: substantiva, adjektiva, verba, adverbia

WordNet

angličtina: PWN (117 tis. synsetů)

WordNet

angličtina: PWN (117 tis. synsetů)

projekty EuroWordNet (angličtina + holandština, italština, španělština, němčina, francouzština, čeština, estonština)

- ILI - InterLingual Index
- Top Ontology (63 kategorií)
- Base Concepts

WordNet

angličtina: PWN (117 tis. synsetů)

projekty EuroWordNet (angličtina + holandština, italština, španělština, němčina, francouzština, čeština, estonština)

- ILI - InterLingual Index
- Top Ontology (63 kategorií)
- Base Concepts

projekty (BalkaNet: bulharština, čeština, rumunština, řečtina, srbština, turečtina), při kterých vznikají wordnety pro další jazyky, koordinátorem databází je Global WordNet Association (GWA)

WordNet

angličtina: PWN (117 tis. synsetů)

projekty EuroWordNet (angličtina + holandština, italština, španělština, němčina, francouzština, čeština, estonština)

- ILI - InterLingual Index
- Top Ontology (63 kategorií)
- Base Concepts

projekty (BalkaNet: bulharština, čeština, rumunština, řečtina, srbština, turečtina), při kterých vznikají wordnety pro další jazyky, koordinátorem databází je Global WordNet Association (GWA)

současný český W.: 28 tis. synsetů


Allen, J. (1995).
Natural Language Understanding (2nd ed.).
Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.


Fellbaum, C. (1998).
WordNet: An Electronic Lexical Database (Language, Speech, and Communication).
The MIT Press.
Published: Hardcover.


Fillmore, C. (1968).
The case for case.
In Bach, E. and Harms, R., editors, *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, New York.


Hlaváčková, D. (2007).
Databáze slovesných valenčních rámců VerbaLex.
PhD thesis, Masarykova univerzita, Filozofická fakulta, Ústav českého jazyka.


Jackendoff, R. (1992).
Semantic Structures.
Current Studies in Linguistics. MIT Press.


Laparra, E. and Rigau, G. (2009).
Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm.
In *RANLP*, Borovets, Bulgaria.


Minsky, M. (1974).
A framework for representing knowledge.
Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA.


Véronis, J. (2004).
Hyperlex: Lexical cartography for information retrieval.
In *Computer Speech and Language: Special Issue on Word Sense Disambiguation*, page 23.