Prezentace pro doktorandský seminář

3. listopadu 2009

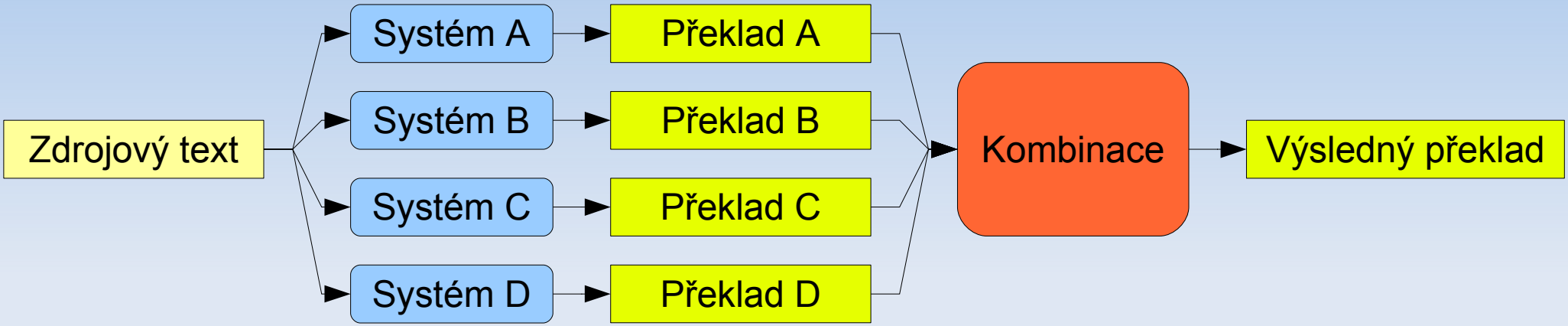
Kombinování překladových systémů

Martin Popel

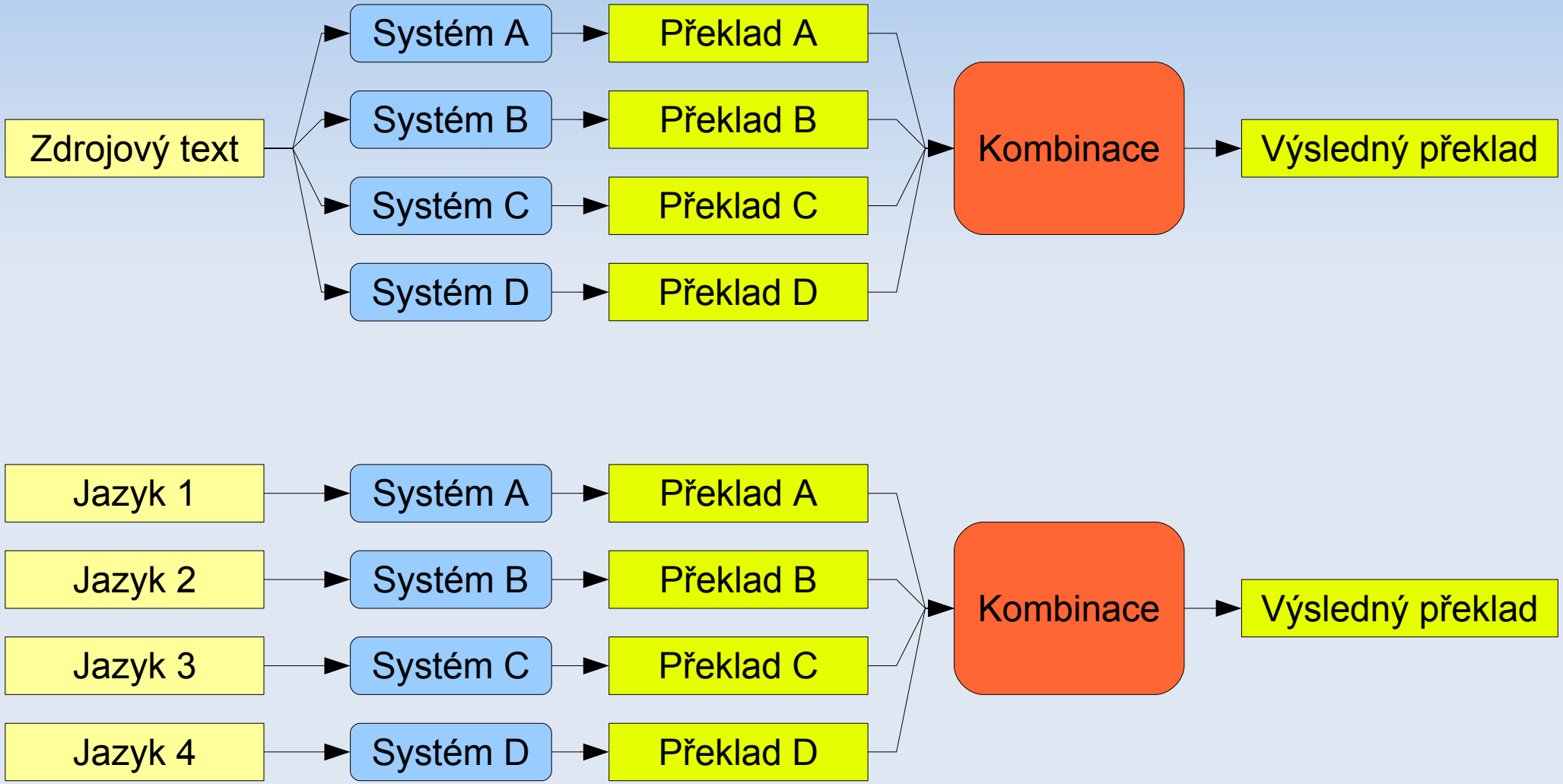
Kombinování překladových (MT) systémů

- Způsoby využití
- Různé způsoby kombinace
- Dva konkrétní přístupy
  - Confusion networks
  - Joint optimization

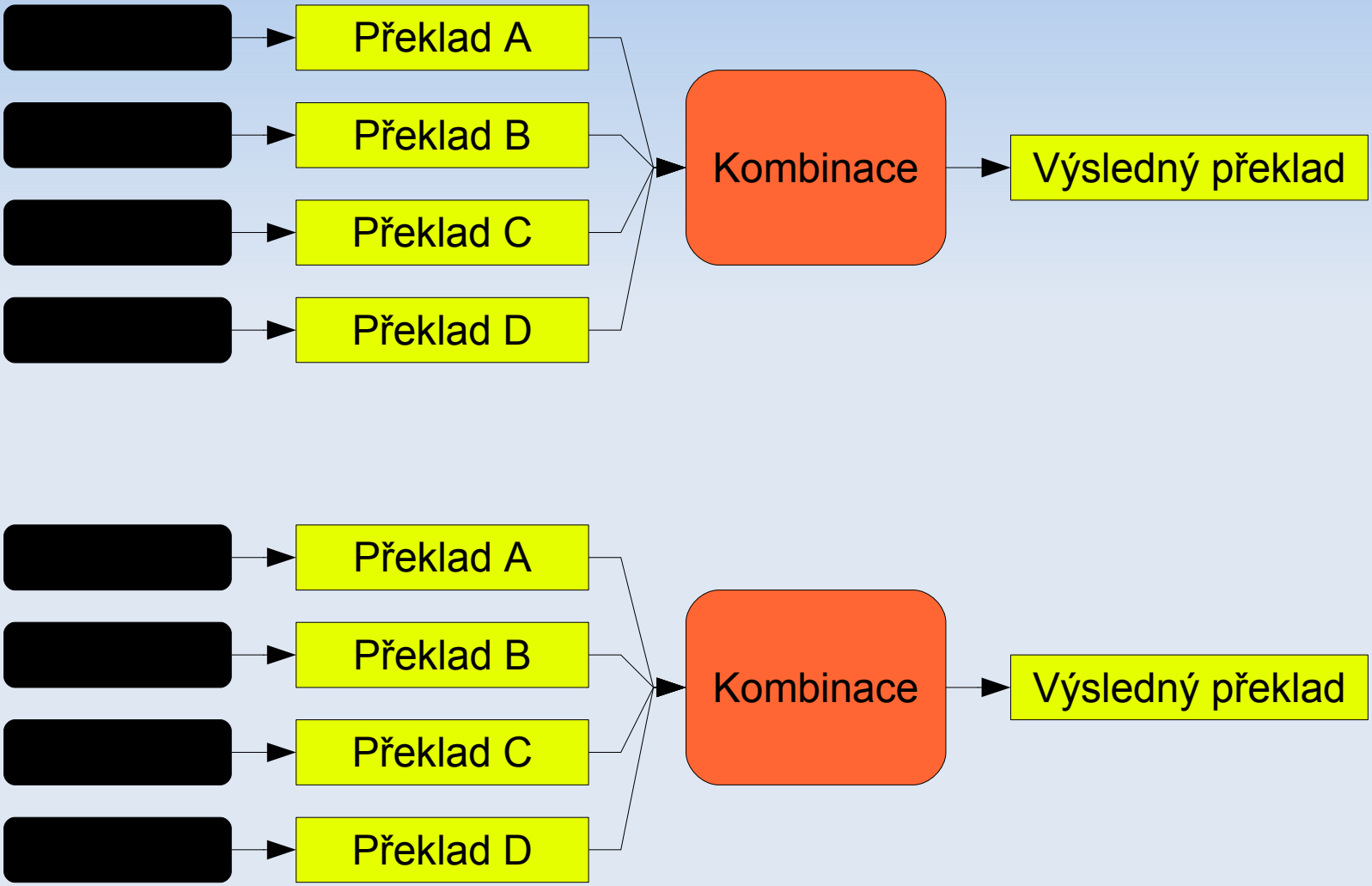
Způsoby využití



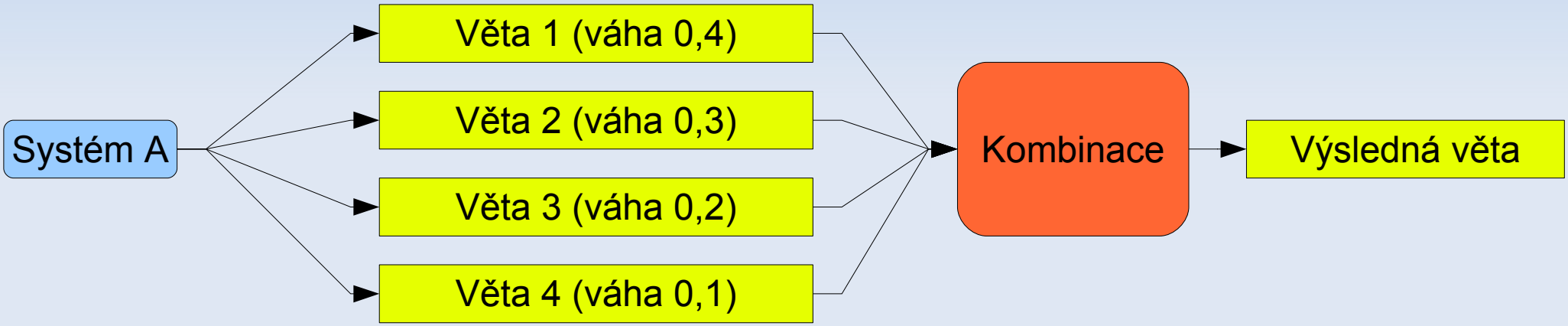
Způsoby využití
jeden zdroj vs. více zdrojů



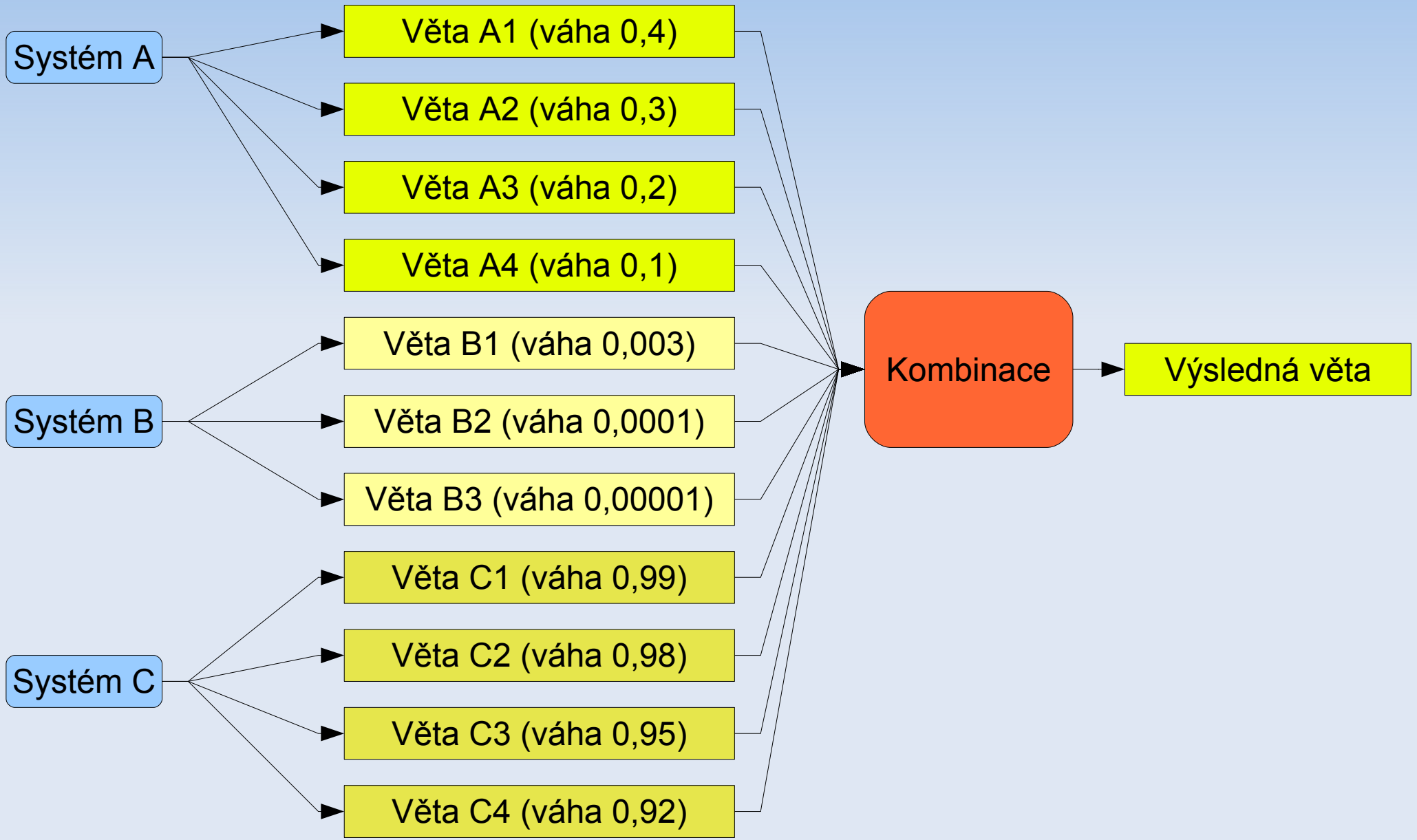
Způsoby využití
jeden zdroj vs. více zdrojů



Způsoby využití
n-best lists



Způsoby využití
obecný případ



Způsoby kombinace
Jak dlouhé segmenty textu?

- Věty

Systém A
Věta A1
Věta A2
Věta A3

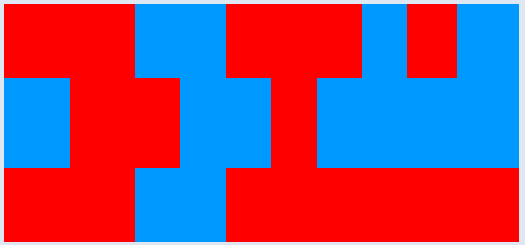
Systém B
Věta B1
Věta B2
Věta B3

Výsledek kombinace
Věta B1
Věta A2
Věta B3

- Fráze

Systém A
Věta A1
Věta A2
Věta A3

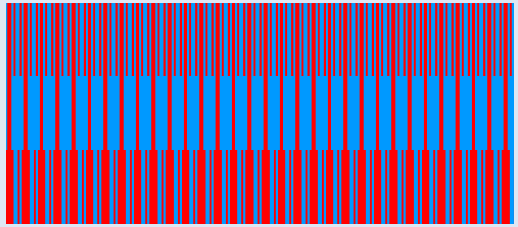
Systém B
Věta B1
Věta B2
Věta B3

Výsledek kombinace


- Slova

Systém A
Věta A1
Věta A2
Věta A3

Systém B
Věta B1
Věta B2
Věta B3

Výsledek kombinace


Způsoby kombinace
Jak dlouhé segmenty textu?

- Věty
  - jednodušší
  - zarovnání po větách se předpokládá už na vstupu
- Fráze a slova
  - Zarovnání po slovech (frázích)
  - Volba slovosledu (pořadí frází)
  - Volba slov (frází)

Způsoby kombinace
Dělení dle Schroeder et al (2009)

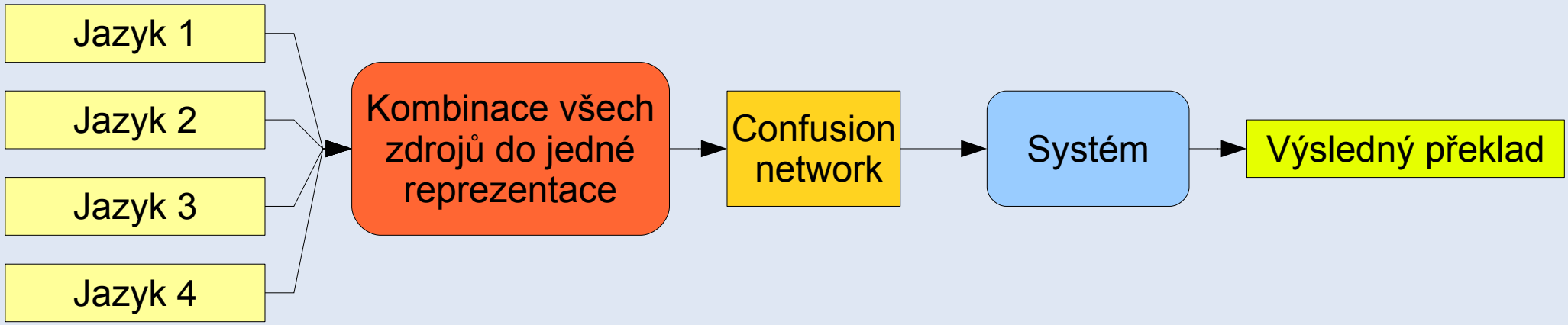
- Věty „Output Selection“
  - jednodušší
  - zarovnání po větách se předpokládá už na vstupu
- Fráze a slova „Output Combination“
  - Zarovnání po slovech (frázích)
  - Volba slovosledu (pořadí frází)
  - Volba slov (frází)

Způsoby kombinace
Dělení dle Schroeder et al (2009)

„Output Selection“

„Input Combination“

„Output Combination“



Způsoby kombinace
Čím se zabývat dál?

„Output Selection“

Malý potenciál !

„Input Combination“

„Output Combination“

Způsoby kombinace
Čím se zabývat dál?

„Output Selection“

Malý potenciál !

„Input Combination“

„Output Combination“

Word-level system combination
outperforms
sentence re-ranking methods.

(Rosti et al, 2007)

Způsoby kombinace
Čím se zabývat dál?

„Output Selection“

Malý potenciál !

„Input Combination“

The potential is high...

(Schroeder et al, 2009)

„Output Combination“

Word-level system combination outperforms sentence re-ranking methods.

(Rosti et al, 2007)

Způsoby kombinace
Čím se zabývat dál?

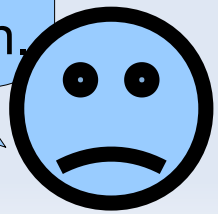
„Output Selection“

Malý potenciál !

„Input Combination“

The potential is high...

… ale výsledky zatím horší než při output combination.



(Schroeder et al, 2009)

„Output Combination“

Word-level system combination outperforms sentence re-ranking methods.

(Rosti et al, 2007)

Output combination
dva přístupy

Confusion networks

1. zarovnání slov
2. volba slovosledu
3. lexikální výběr

Word-level system combination based on confusion networks outperforms sentence re-ranking methods.

Joint optimization

Output combination
dva přístupy

Confusion networks

1. zarovnání slov
2. volba slovosledu
3. lexikální výběr

Word-level system combination based on confusion networks outperforms sentence re-ranking methods.

Joint optimization
(He and Toutanova, 2009)

1. + 2. + 3.

Joint optimization approach significantly outperforms confusion-network-based systems.

Confusion networks

1. zarovnání slov 2. volba slovosledu 3. lexikální výběr

h_1 This is a dog. váha(1) = 0,4

h_2 It is dog of mine. váha(2) = 0,3

h_3 This are my dog. váha(3) = 0,2

h_4 It is our dog. váha(4) = 0,1

$\mathbf{H} = \{h_1, h_2, h_3, h_4\}$

$h_1 = \{h_{1,1}, h_{1,2}, h_{1,3}, h_{1,4}\} = \{\text{this, is, a, dog}\}$

$h_1 = \{h_{2,1}, h_{2,2}, h_{2,3}, h_{2,4}, h_{2,5}\} = \{\text{it, is, dog, of, mine}\}$

Confusion networks

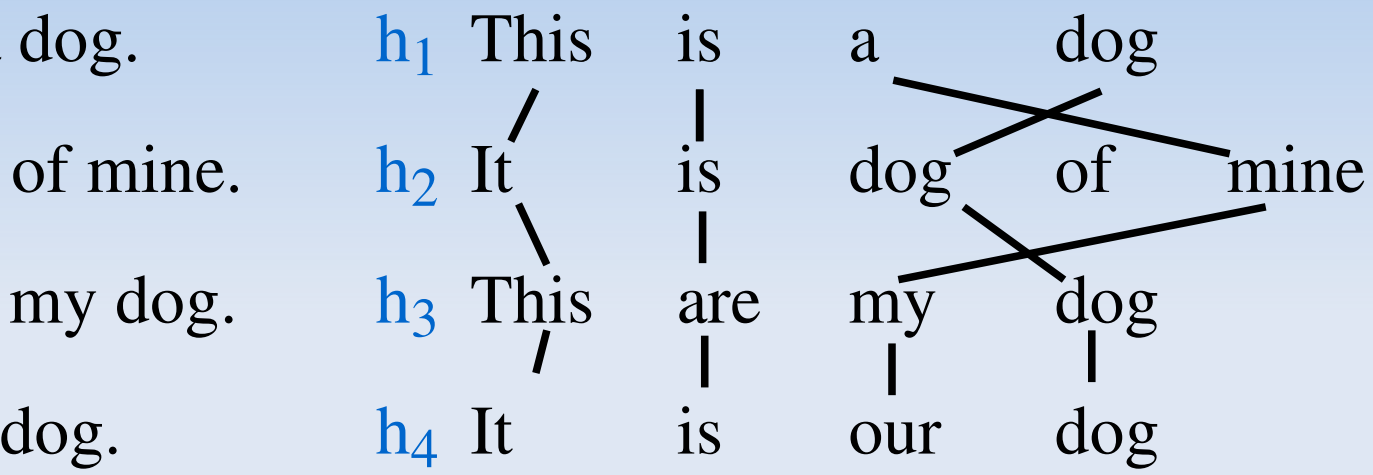
1. zarovnání slov 2. volba slovosledu 3. lexikální výběr

h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

h_4 It is our dog.



Confusion networks

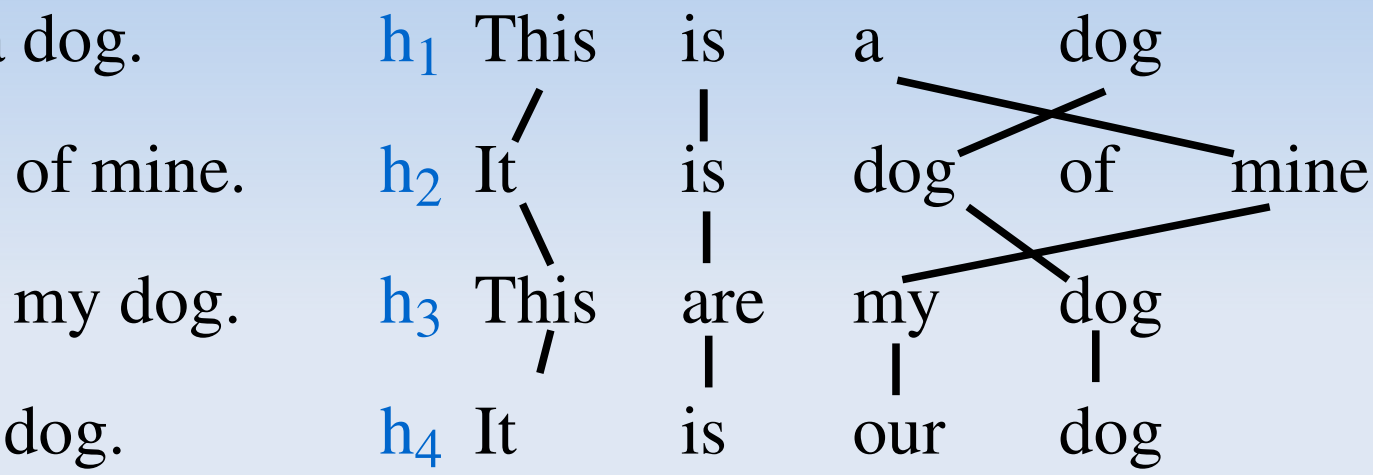
1. zarovnání slov 2. volba slovosledu 3. lexikální výběr

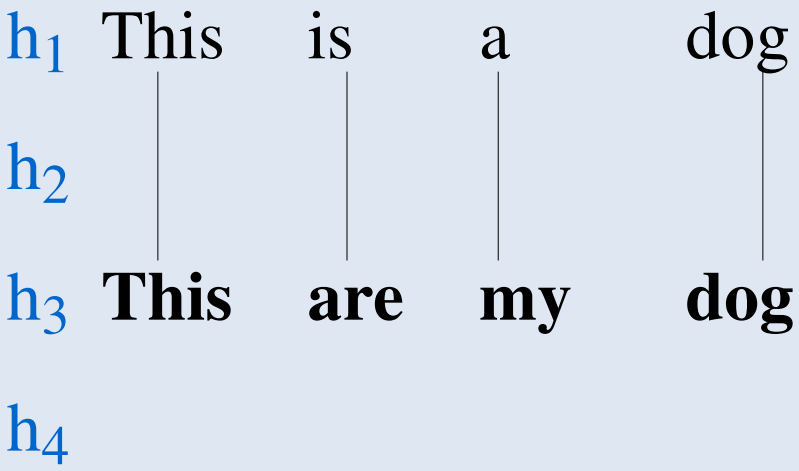
h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

h_4 It is our dog.





Confusion networks

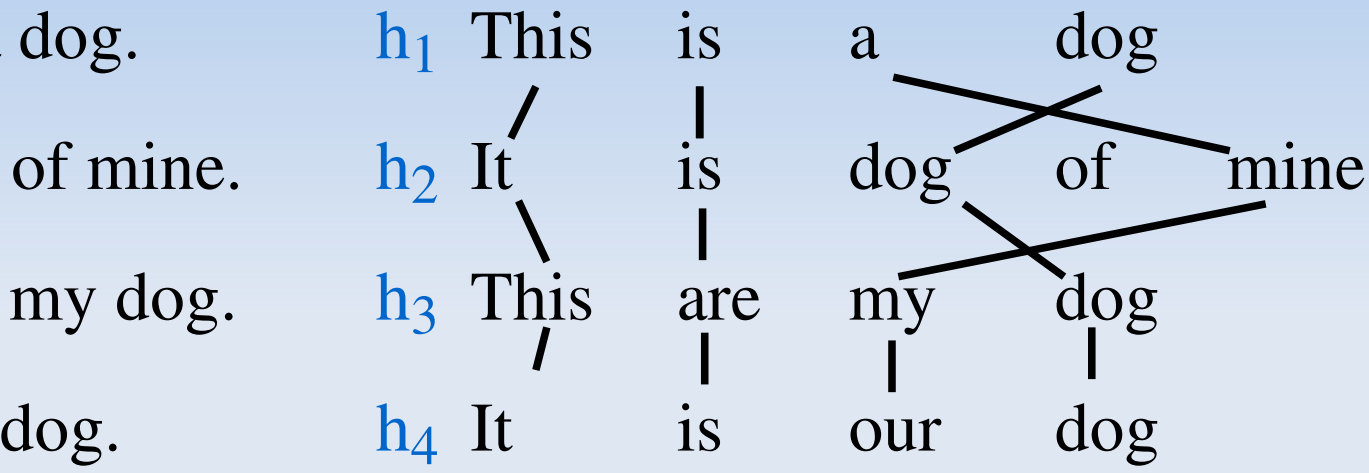
1. zarovnání slov 2. volba slovosledu 3. lexikální výběr

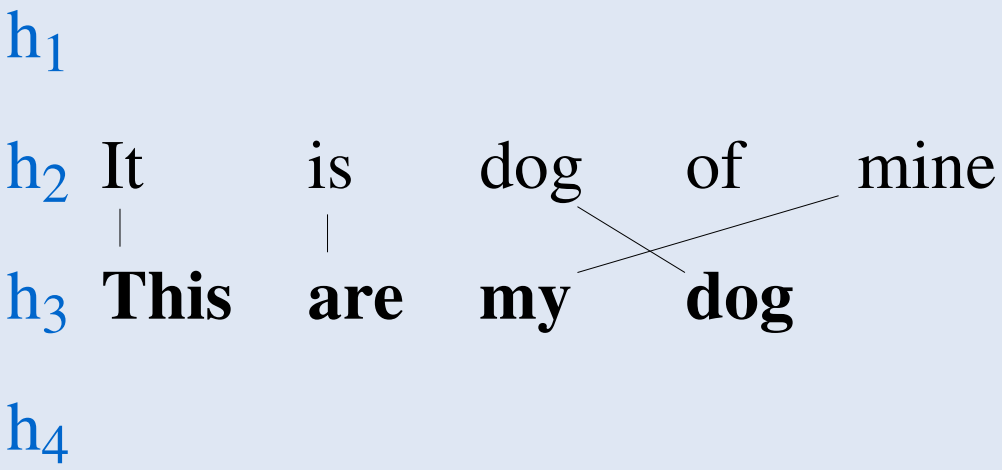
h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

h_4 It is our dog.





Confusion networks

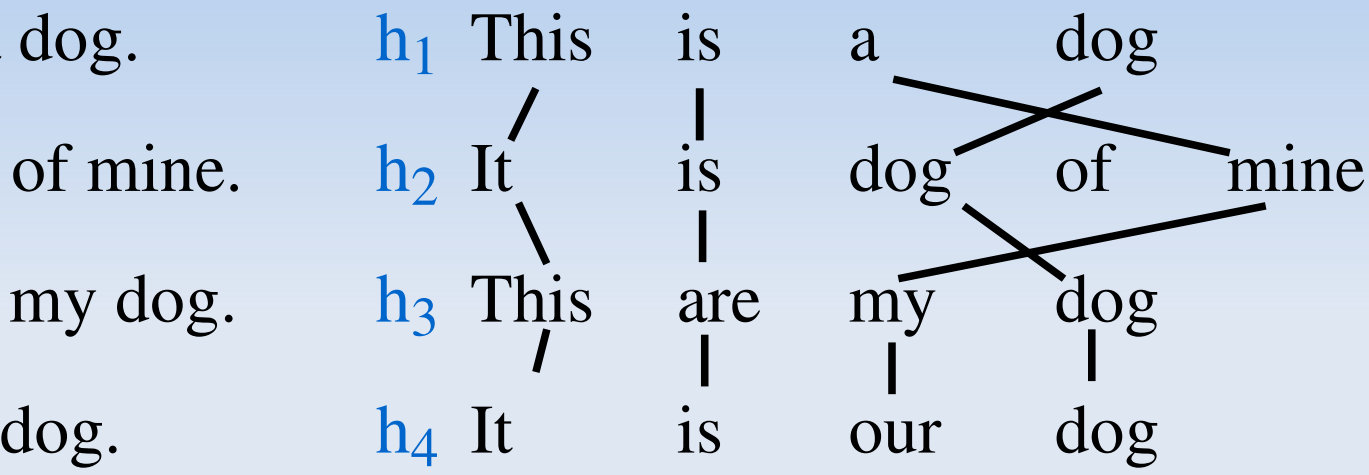
1. zarovnání slov 2. volba slovosledu 3. lexikální výběr

h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

h_4 It is our dog.



h_1

h_2

h_3 **This** **are** **my** **dog**

h_4 It is our dog

Confusion networks

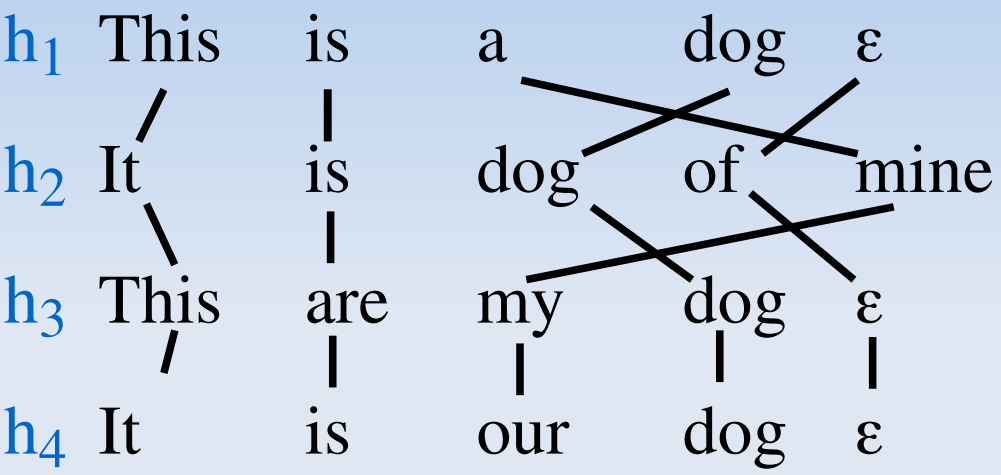
1. zarovnání slov 2. volba slovosledu 3. lexikální výběr

h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

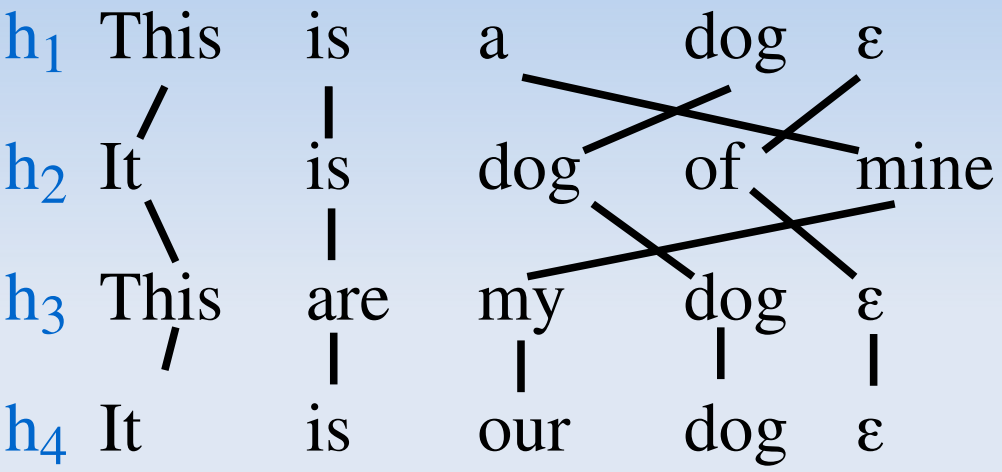
h_4 It is our dog.



Confusion networks

1. zarovnání slov **2. volba slovosledu** 3. lexikální výběr

h_1 This is a dog.
h_2 It is dog of mine.
h_3 This are my dog.
h_4 It is our dog.



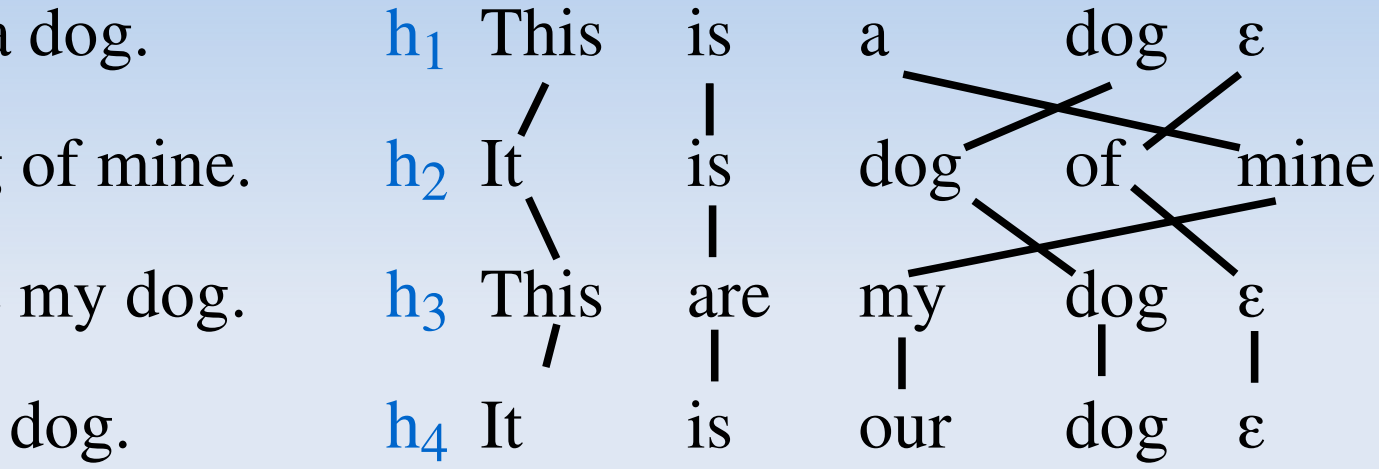
CS = correspondence set

	CS_1	CS_2	CS_3	CS_4	CS_5
h_1	This	is	a	dog	ε
h_2	It	is	mine	dog	of
h_3	This	are	my	dog	ε
h_4	It	is	our	dog	ε

Confusion networks

1. zarovnání slov **2. volba slovosledu** 3. lexikální výběr

h_1 This is a dog.
h_2 It is dog of mine.
h_3 This are my dog.
h_4 It is our dog.



C = {CS_1, CS_2, CS_3, CS_4, CS_5}

CS_1 = CS(1, 1, 1, 1)

CS_2 = CS(2, 2, 2, 2)

CS_3 = CS(3, 5, 3, 3)

CS_4 = CS(4, 3, 4, 4)

CS_5 = CS(0, 4, 0, 0)

	CS_1	CS_2	CS_3	CS_4	CS_5
h_1	This	is	a	dog	ε
h_2	It	is	mine	dog	of
h_3	This	are	my	dog	ε
h_4	It	is	our	dog	ε

Confusion networks

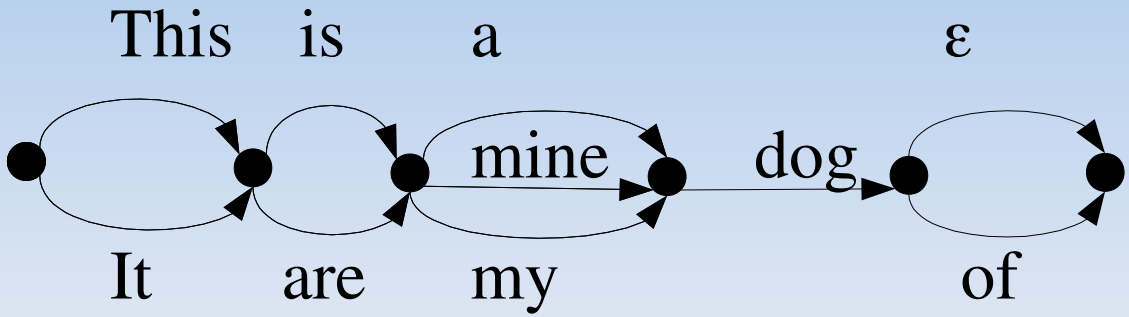
1. zarovnání slov **2. volba slovosledu** 3. lexikální výběr

h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

h_4 It is our dog.



↑ confusion network ↑
↓ corespondence sets ↓

	CS_1	CS_2	CS_3	CS_4	CS_5
h_1	This	is	a	dog	ε
h_2	It	is	mine	dog	of
h_3	This	are	my	dog	ε
h_4	It	is	our	dog	ε

Confusion networks

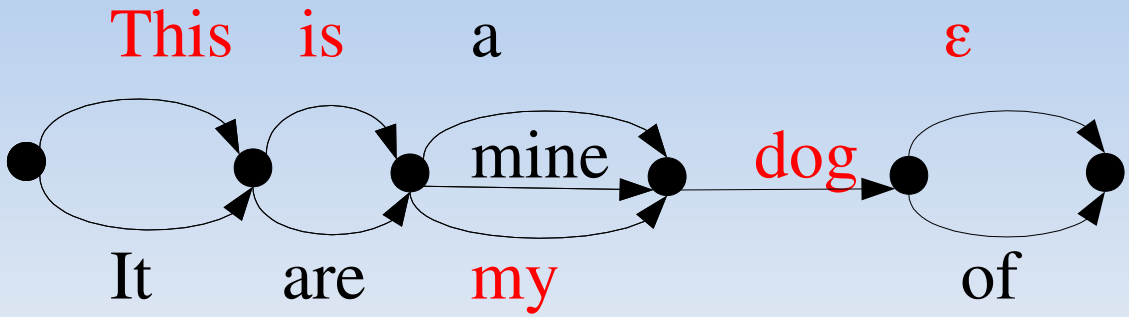
1. zarovnání slov
2. volba slovosledu
3. lexikální výběr

h_1 This is a dog.

h_2 It is dog of mine.

h_3 This are my dog.

h_4 It is our dog.



↑ confusion network ↑
↓ corespondence sets ↓

	CS_1	CS_2	CS_3	CS_4	CS_5
h_1	This	is	a	dog	ε
h_2	It	is	mine	dog	of
h_3	This	are	my	dog	ε
h_4	It	is	our	dog	ε

Joint optimization

Log-lineární model

$$best = \underset{w,O,C}{argmax} \sum \alpha_i \cdot f_i(w, O, C, H)$$

w – posloupnost slov

C – množina CS

O – pořadí CS

H – množina vstupních hypotéz

α – vektor vah

f – features

Joint optimization features

- Tri-gram language model
- Bi-gram voting model
- Word posterior model
- Distortion model
- Alignment model
- a další (počet slov, počet CS)

Otevřené otázky

- Využití faktorizovaných modelů
  - např.

Systém A
Lepší lexikální výběr

Gramatická shoda dodržena pouze u častých n-gramů

Systém B
Horší lexikální výběr

Gramatické věty

- Odolnost vůči přidání méně kvalitních překladových systémů
  - Jak trénovat váhy systémů, aby přidání dalších nikdy neuškodilo?

Shrnutí

- **Využití**: více systémů, více zdrojů, n-best lists
- **Druhy kombinací**
  - po větách, po frázích, po slovech
  - použitelné:
    - pro jakékoli systémy (black-box MT)
    - šité na míru konkrétnímu MT systému
- **Confusion network** přístup
  - 1. zarovnání slov, 2. slovosled, 3. lexikální výběr
- **Joint optimization** přístup
  - kroky 1.,2.,3. prováděny naráz, jeden log-lineární model

Literatura

- Josh Shroeder, Trevor Cohn, Philipp Koehn:
  Word Lattices for Multi-Source Translation
  Proceedings of EACL, 2009

- Xiadong He, Kristina Toutanova:
  Joint Optimization for Machine Translation System Combination
  Proceedings of EMNLP, 2009

- Antti-Veikko I. Rosti et al.:
  Combining Outputs from Multiple Machine Translation Systems
  Proceedings of NAACL, 2007