Jaká data se používají ve strojovém překladu

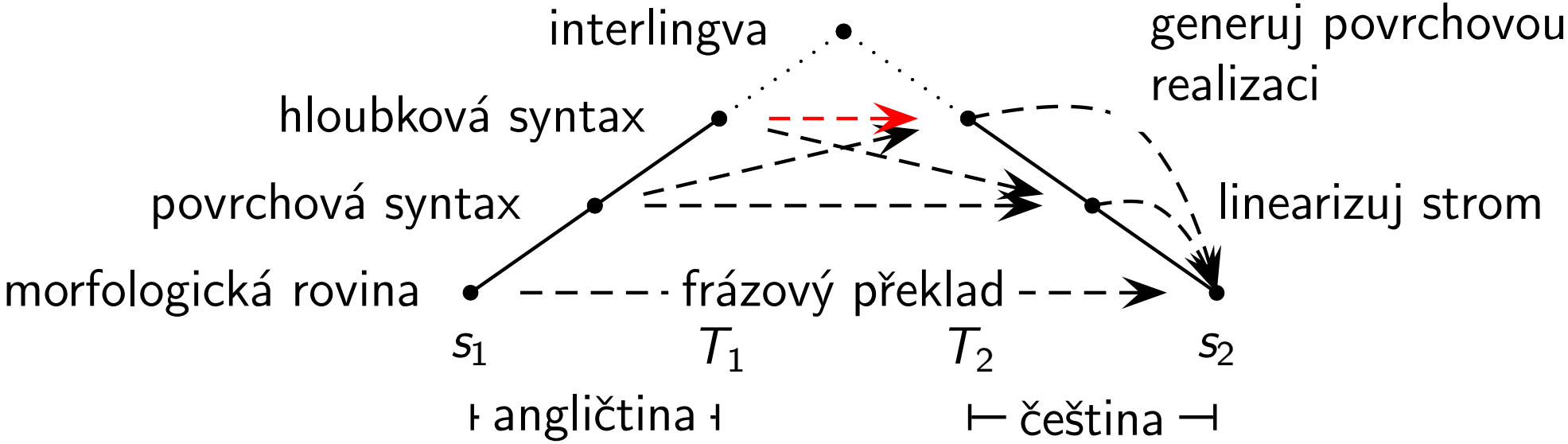
Ondřej Bojar
bojar@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
MFF UK

ELRC Workshop, 14. duben 2016

Osnova

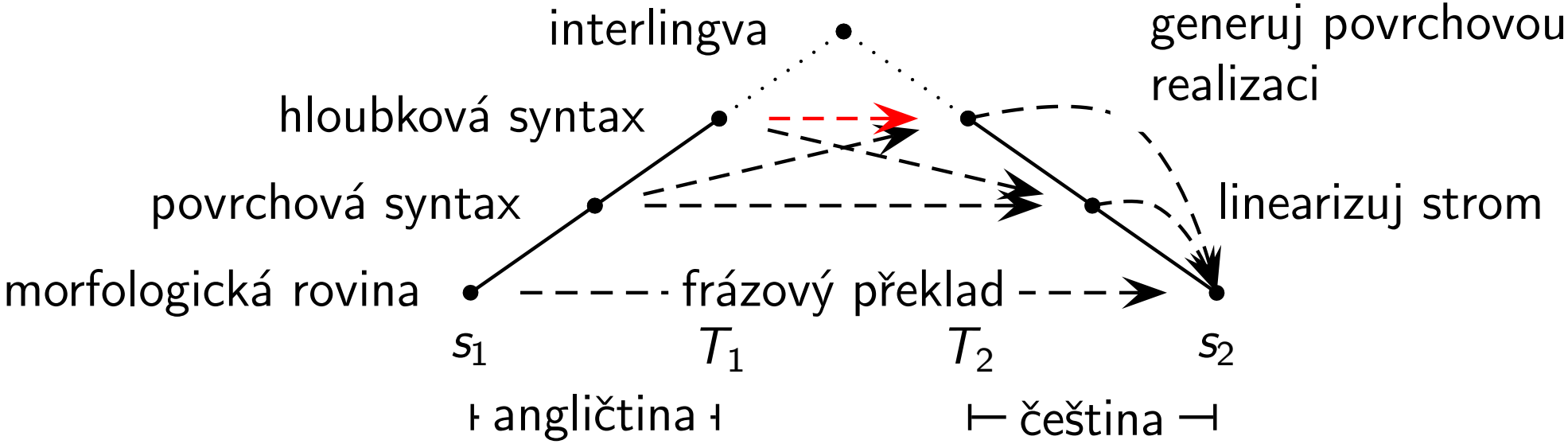
- Typy dat ve strojovém překladu.
- Kolik je potřeba?
- Absolutní minimum: paralelní korpusy.
  - Získávání paralelních textů z webu.
- Sdílení zdrojů.

Přístupy ke strojovému překladu



- Pravidlový vs. statistický přístup:
  - Pravidlové systémy píší lingvisté-programátoři.
  - Statistické systémy se naučí samy podle dat.

Přístupy ke strojovému překladu



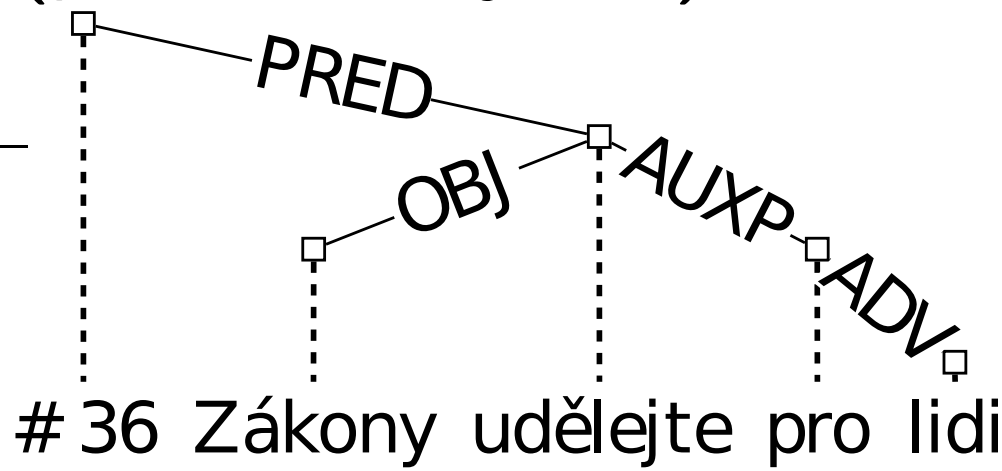
- Pravidlový vs. statistický přístup:
  - ~~Pravidlové systémy píší lingvisté-programátoři.~~
  - Statistické systémy se naučí samy podle dat.

Formální popis češtiny ⇝ Ruční anotace

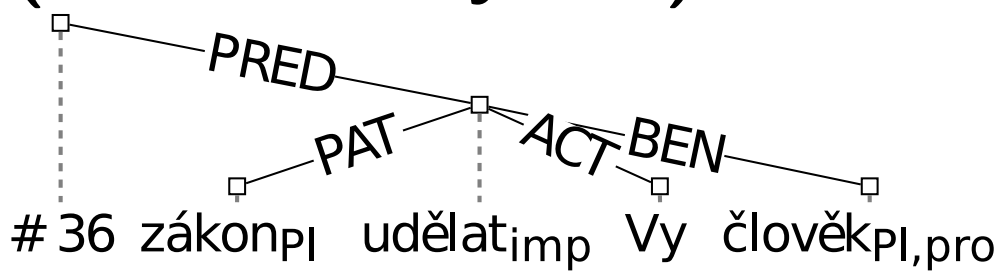
Morfologická rovina:

Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Analytická rovina (povrchová syntax):



Tektogramatická rovina (hloubková syntax):



Lingvistická data

- **Korpusy** jsou (velké) sbírky textů:
  - Texty typicky označkované nebo včetně větných rozborů.
    Pražský závislostní korpus (PDT): 1.5 mil. slov.
    Pražský čj-aj závislostní korpus (PCEDT): 50 tis. vět.
  - Některé vícejazyčné: CzEng (15 mil. vět, 220 mil. slov, odpovídá ∼50 metrům knih, ty tvoří však jen čtvrtinu).
- **Slovníky** strojově čitelné:
  - Morfologický slovník říká, že *kočka* je české slovo a *kočke* ne.
  - Valenční slovník říká, že:

    Rodiče přijali Petra. $\rightarrow$ je správně

    Rodiče přijeli Petra. $\rightarrow$ není správně
  - Slovník subjektivity obsahuje hodnotící výrazy.

$\Rightarrow$ Lze využít v programech (pravidlových i statistických).

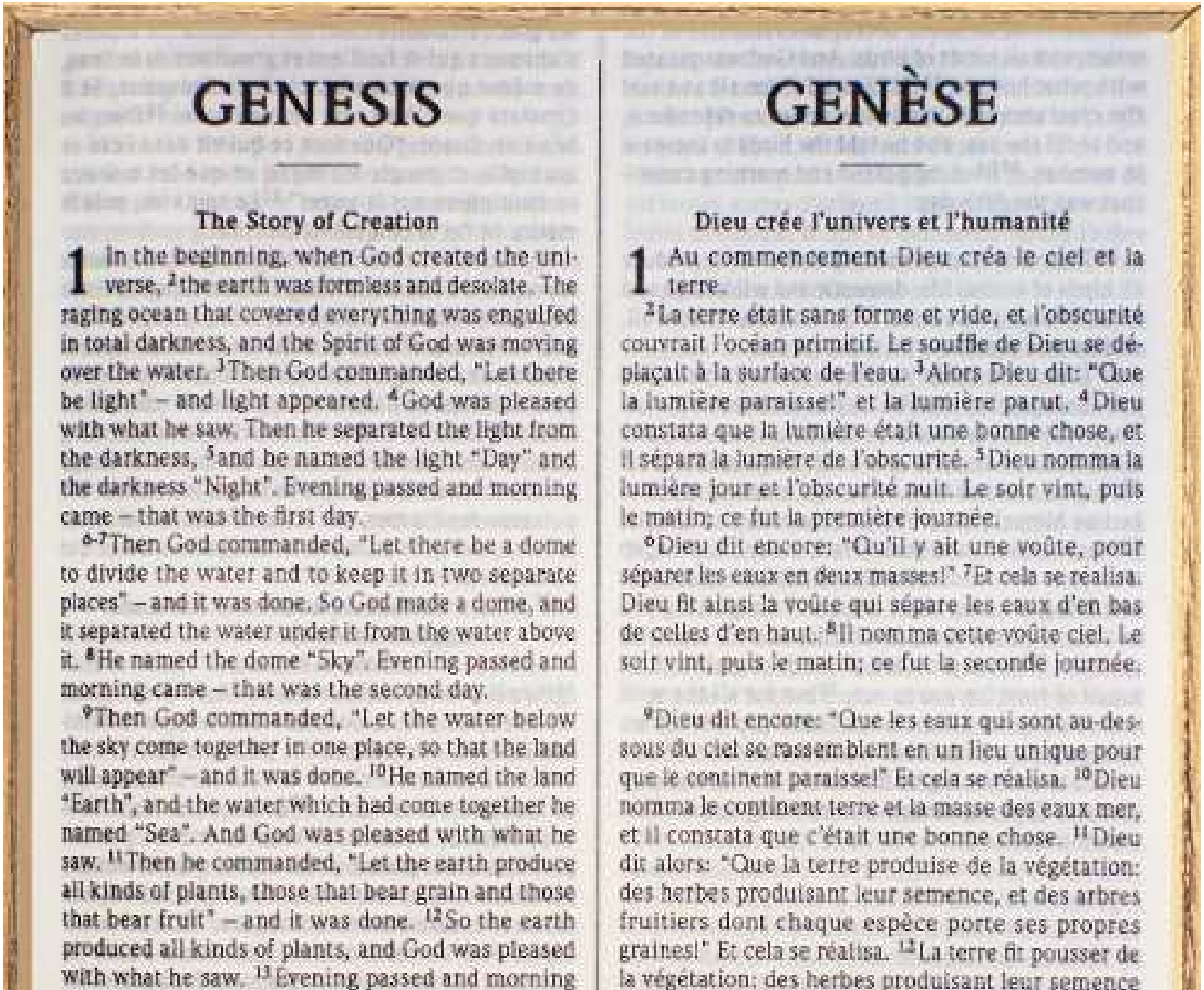
Data po frázový překlad (Moses, MT@EC)

- Překladové slovníky méně vhodné.
  - Překlad potřebuje vidět slova v kontextu.
- Terminologie a glosáře méně vhodné.
  - Současné metody mohou vynutit překlad jen velice tupě.

- Zásadní je paralelní korpus.
- Pro kvalitu důležité, aby texty byly z dané domény.

- Systém memoruje až desetice slov.
  - $\Rightarrow$ Dlouhé termíny zachovány přesně.
  - $\Rightarrow$ České tvarosloví a volný slovosled situaci komplikuje.

Paralelní korpus



GENESIS

The Story of Creation

1 In the beginning, when God created the universe, 2 the earth was formless and desolate. The raging ocean that covered everything was engulfed in total darkness, and the Spirit of God was moving over the water. 3 Then God commanded, "Let there be light" – and light appeared. 4 God was pleased with what he saw. Then he separated the light from the darkness, 5 and he named the light "Day" and the darkness "Night". Evening passed and morning came – that was the first day.

6-7 Then God commanded, "Let there be a dome to divide the water and to keep it in two separate places" – and it was done. So God made a dome, and it separated the water under it from the water above it. 8 He named the dome "Sky". Evening passed and morning came – that was the second day.

9 Then God commanded, "Let the water below the sky come together in one place, so that the land will appear" – and it was done. 10 He named the land "Earth", and the water which had come together he named "Sea". And God was pleased with what he saw. 11 Then he commanded, "Let the earth produce all kinds of plants, those that bear grain and those that bear fruit" – and it was done. 12 So the earth produced all kinds of plants, and God was pleased with what he saw. 13 Evening passed and morning

GENÈSE

Dieu crée l'univers et l'humanité

1 Au commencement Dieu créa le ciel et la terre.

2 La terre était sans forme et vide, et l'obscurité couvrait l'océan primitif. Le souffle de Dieu se déplaçait à la surface de l'eau. 3 Alors Dieu dit: "Que la lumière paraisse!" et la lumière parut. 4 Dieu constata que la lumière était une bonne chose, et il sépara la lumière de l'obscurité. 5 Dieu nomma la lumière jour et l'obscurité nuit. Le soir vint, puis le matin; ce fut la première journée.

6 Dieu dit encore: "Qu'il y ait une voûte, pour séparer les eaux en deux masses!" 7 Et cela se réalisa. Dieu fit ainsi la voûte qui sépare les eaux d'en bas de celles d'en haut. 8 Il nomma cette voûte ciel. Le soir vint, puis le matin; ce fut la seconde journée.

9 Dieu dit encore: "Que les eaux qui sont au-dessous du ciel se rassemblent en un lieu unique pour que le continent paraisse!" Et cela se réalisa. 10 Dieu nomma le continent terre et la masse des eaux mer, et il constata que c'était une bonne chose. 11 Dieu dit alors: "Que la terre produise de la végétation: des herbes produisant leur semence, et des arbres fruitiers dont chaque espèce porte ses propres graines!" Et cela se réalisa. 12 La terre fit pousser de la végétation: des herbes produisant leur semence

⇝ Automatické zarovnání po větách

In my dream , there was a sycamore growing out of the ruins of the sacristy , and I was told that , if I dug at the roots of the sycamore , I would find a hidden treasure . But I ' m not so stupid as to cross an entire desert just because of a recurrent dream . " And they disappeared . The boy stood up shakily , and looked once more at the Pyramids . " It is I who dared to do so , " said the boy . This man looked exactly the same , except that now the roles were reversed . " It is I who dared to do so , " he

अपने सपने में मुझे एक गुलर का पेड़ दिखाई देता था और मुझे लगता था कि अगर मैं उस गुलर की जड़ें खोद डालूं तो मुझे छिपा हुआ खजाना मिल जाएगा । मगर मैं तुम्हारी तरह इतना बेवकूफ नहीं हूं कि महज बार – बार आने वाले एक सपने के कारण पूरे रेगिस्तान को पार करूं । वे लोग , उसके बाद वहां से चले गए । लड़का लड़खड़ाता हुआ किसी तरह खड़ा हो गया ।<s>एक बार फिर उसने पिरामिडों को देखा । " यह जुर्रत मैंने की थी , " लड़के ने कहा ।<s>उसे सेंटियागो मातामोरोस की वह प्रतिमा याद आई जिसमें वह घोड़े पर सवार था और उसके घोड़े के खुरों में कितने ही नास्तिक कुचले हुए पड़े थे । यह घुड़सवार भी बिलकुल वैसा ही था । यह बात और थी कि इनके किरदार बदले हुए थे । " मैंने ही ऐसा करने का साहस किया था , " लड़के ने दोहराया और अपनी गर्दन तलवार का वार सहने के लिए झुका दी । ' जिंदगी ने भी हमेशा मेरे साथ अच्छा बर्ताव किया । '

⇝ Automatické zarovnání po větách

In my dream , there was a sycamore growing out of the ruins of the sacristy , and I was told that , if I dug at the roots of the sycamore , I would find a hidden treasure .	अपने सपने में मुझे एक गुलर का पेड़ दिखाई देता था और मुझे लगता था कि अगर मैं उस गुलर की जड़ें खोद डालूं तो मुझे छिपा हुआ खजाना मिल जाएगा ।
But I ' m not so stupid as to cross an entire desert just because of a recurrent dream .	" मगर मैं तुम्हारी तरह इतना बेवकूफ नहीं हूं कि महज बार - बार आने वाले एक सपने के कारण पूरे रेगिस्तान को पार करूं ।
And they disappeared .	वे लोग , उसके बाद वहां से चले गए ।
The boy stood up shakily , and looked once more at the Pyramids .	लड़का लड़खड़ाता हुआ किसी तरह खड़ा हो गया । एक बार फिर उसने पिरामिडों को देखा ।
	" यह जुर्रत मैंने की थी , " लड़के ने कहा । उसे सेंटियागो मातामोरोस की वह प्रतिमा याद आई जिसमें वह घोड़े पर सवार था और उसके घोड़े के खुरों में कितने ही नास्तिक कुचले हुए पड़े थे ।
" It is I who dared to do so , " said the boy .	यह घुड़सवार भी बिलकुल वैसा ही था ।
This man looked exactly the same , except that now the roles were reversed .	यह बात और थी कि इनके किरदार बदले हुए थे ।
" It is I who dared to do so , " he repeated , and he lowered his head to receive a blow from the sword .	" मैंने ही ऐसा करने का साहस किया था , " लड़के ने दोहराया और अपनी गर्दन तलवार का वार सहने के लिए झुका दी ।
" Life was good to me , " the man said .	' जिंदगी ने भी हमेशा मेरे साथ अच्छा बर्ताव किया । '
" When you appeared in my dream , I felt that all my efforts had been rewarded , because my son ' s poems will be read by men for generations to come .	उस आदमी ने कहा , ' जब आप मेरे सपने में आए थे , तो मुझे लगा कि मैंने अपने कर्मों का पुरस्कार पा लिया ...
	मेरे लिए इससे बढ़कर और क्या बात होती कि मेरे बेटे की कविताएं युग - युगों तक पढी जाएं ।
I don ' t want anything for myself .	नहीं , मुझे अपने लिए कुछ नहीं चाहिए ।
But any father would be proud of the fame achieved by one whom he had cared for as a child , and educated as he grew up .	कोई भी बाप उस इंसान की शोहरत सुनकर फूला नहीं समाएगा जिसे उसने अपनी गोद में खिलाया , पढ़ाया - लिखाया और पाल - पोसकर बड़ा किया हो ।
" We ' re two very different things .	" " हम दो अलग - अलग चीजें हैं । "
" That ' s not true , " the boy said .	" यह सही नहीं है । " लड़के ने कहा ,
" I learned the alchemist ' s secrets in my travels .	" यात्रा के दौरान मैंने कीमियागर के रहस्यों को जाना है ।
I have inside me the winds , the deserts , the oceans , the stars , and everything created in the universe .	मेरे ही भीतर सब छिपा है — हवा , रेगिस्तान , समुद्र , तारे और वह सब कुछ जो ब्रह्माण्ड ने सर्जित किया है ।
We were all made by the same hand , and we have the same soul .	हम सबको उसी हाथ ने बनाया और हम सबकी आत्मा भी एक ही है ।
You ' ll learn to love the desert , and you ' ll get to know every one of the fifty thousand palms .	तुम्हें रेगिस्तान से प्यार करना आ जाएगा और उन पचास हजार खजूर के पेड़ों में तुम एक - एक को पहचानने लगोगे ।
You ' ll watch them as they grow , demonstrating how the world is always changing .	उन्हें बढ़ता हुआ देखकर तुम अनुभव करोगे कि कैसे हर क्षण दुनिया बदलती रहती है ।
And you ' ll get better and better at understanding omens , because the desert is the best teacher there is .	तुम शकुन पहचानने में बेहतर से बेहतर बनते जाओगे चूंकि इस मामले में रेगिस्तान से बढ़कर कोई अच्छा गुरु नहीं है ।
" Sometime during the second year , you ' ll remember about the treasure .	" फिर , किसी वक्त , दूसरे साल के दौरान तुम्हें खजाने की याद सताएगी ।
The omens will begin insistently to speak of it , and you ' ll try to ignore them .	शकुन फौरन तुम्हें उसके बारे में बताना शुरू कर देंगे , मगर तुम उन्हें अनदेखा करना चाहोगे ।
But you know that I ' m not going to go to Mecca . Just as you know that you ' re not going to buy your sheep .	" तुम अच्छी तरह से जानते हो , कि मैं मक्का नहीं जाने वाला हूं ठीक उसी तरह जैसे कि तुम कोई भेड़ - बेड़ नहीं खरीदने वाले हो ! "
" Who told you that ? " asked the boy , startled .	" आपसे ऐसा किसने कहा ? " लड़के को आश्चर्य हुआ ।
" Maktub " said the old crystal merchant .	" मकतूब ! " क्रिस्टल - व्यापारी ने कहा ,
And he gave the boy his blessing .	कुछ पल खामोश रह कर , उसने लड़के को भरपूर आशीर्वाद दिया ।
The boy went to his room and packed his belongings .	कमरे में जाकर लड़के ने अपना सामान बांधा ।
They filled three sacks .	तीन बोरे भर गए ।
As he was leaving , he saw , in the corner of the room , his old shepherd ' s pouch .	बाहर जाते हुए उसने कमरे के एक कोने में , अपनी पुरानी थैली देखी ।
" I want to see the greatness of Allah , " the chief said , with respect .	" मैं अल्लाह की महानता देखना चाहता हूं । " बड़े आदर के साथ मुखिया ने कहा ।
" I want to see how a man turns himself into the wind .	" " मैं देखना चाहता हूं कि कैसे कोई आदमी खुद को हवा में बदलता है । "
But he made a mental note of the names of the two men who had expressed their fear .	मगर उसने अपने मन में उन दो सेनापतियों के नाम याद कर लिए जिन्होंने डर का इजहार किया था ।
As soon as the wind stopped , he was going to remove them from their commands , because true men of the desert are not afraid .	जैसे ही यह अंधड़ रुकेगा , वह उन दोनों को सेना - कमान से हटा देगा क्योंकि रेगिस्तान के असली निवासी कभी किसी बात से डरते नहीं ।
" Tell them about the armies that are approaching .	" " युद्ध के लिए इधर आती हुई सेनाओं की सूचना उन्हें दो । "
" They ' ll laugh at me .	" " वे मुझ पर हंसेंगे । "
" They are men of the desert , and the men of the desert are used to dealing with omens .	" " वे रेगिस्तान के लोग हैं । शकुनों पर विश्वास करते हैं । "
" Well , then , they probably already know .	" " तब तो उन्हें मालूम चल ही गया होगा । "
" They ' re not concerned with that right now .	" अभी उनका ध्यान इस तरफ नहीं है ।
They believe that if they have to know about something Allah wants them to know , someone will tell them about it .	उनका विश्वास है कि अगर अल्लाह की यह मर्जी है कि उन्हें कुछ पता चले तो कोई न कोई आकर उन्हें उसके बारे में बता देगा ।
It has happened many times before .	ऐसा पहले भी कई बार हो चुका है ।
" What about one-tenth of my treasure ?	" " बजाय भेड़ों के क्या खजाने का दसवीं भाग नहीं चलेगा ? "
The old man looked disappointed .	बूढ़ा निराश होते हुए बोला ,
" If you start out by promising what you don ' t even have yet , you ' ll lose your desire to work toward getting it .	" " अगर तुम उस वस्तु को देने का वादा करते हो जो तुम्हारे पास अभी है ही नहीं , तो तुम उसे पाने के लिए प्रयत्न करने की इच्छा भी छोड़ सकते हो । "
The boy told him that he had already promised to give one-tenth of his treasure to the Gypsy .	लड़के ने बताया कि खजाने का दसवीं हिस्सा देने का वादा उसने जिप्सी महिला से कर डाला था ।
" Gypsies are experts at getting people to do that , " sighed the old man .	बूढ़े ने लंबी सांस लेते हुए कहा , „ जिप्सी लोग ऐसा वादा लेने में बड़े चतुर होते हैं ।
The elder continued , " When the pharaoh dreamed of cows that were thin and cows that were fat , this man I ' m speaking of rescued Egypt from famine .	
	वृद्ध ने अपनी बात जारी रखी , " जब फैरों ने सपने में ऐसी गायें देखी जिनमें से कुछ बहुत दुबला थीं और कुछ बहुत मोटी , तो उस आदमी से जिसका मैंने अभी जिक्र किया — पूरे मिस्र को अकाल से बचाया था ।
His name was Joseph .	उसका नाम था जोसेफ ।
He , too , was a stranger in a strange land , like you , and he was probably about your age .	" वह भी एक अजनबी देश में एक अजनबी ही था , ठीक तुम्हारी ही तरह ... लगभग तुम्हारी ही उम्र का । "
He paused , and his eyes were still unfriendly .	वह चुप हो गया मगर आंखों में अभी भी दोस्ती की कोई झलक नहीं थी ।
" We always observe the Tradition .	" हम हमेशा परंपराओं का पालन करते हैं ।
" Everyone on earth has a treasure that awaits him , " his heart said .	हर किसी के लिए इस धरती पर एक खजाना है जो उसका इंतजार कर रहा है ।
" We , people ' s hearts , seldom say much about those treasures , because people no longer want to go in search of them .	दिल ने कहा , ' मगर हम इन खजानों के बारे में शायद ही कभी बात करते हैं ... क्योंकि लोग अब उन्हें खोजना ही नहीं चाहते ।
We speak of them only to children . Later , we simply let life proceed , in its own direction , toward its own fate .	हम उनकी बात केवल बच्चों से करते हैं और फिर जीवन जैसा चाहे उस दिशा में बह जाए - यह उसका भाग्य ।
But he didn ' t need to worry about that right now .	लेकिन अभी के लिए चिंता का यह विषय नहीं था ।
What he had to be concerned about was his treasure , and how he was going to go about getting it .	चिंता तो उसे खजाने की करनी थी और उस तक कैसे पहुंचा जाए , इसकी ।
The sale of his sheep had left him with enough money in his pouch , and the boy knew that in money there was magic ;	भेड़ों की बिक्री से उसके पास काफी पैसा आ गया था । वह जानता था पैसे में बड़ा जादू होता है । पैसा हो तो आदमी कभी अकेला नहीं रहता ।
Before long , maybe in just a few days , he would be at the Pyramids .	थोड़े ही दिनों में वह पिरामिडों में होगा ।
An old man , with a breastplate of gold , wouldn ' t have lied just to acquire six sheep .	रत्नजटित स्वर्ण कवचवाला वह वृद्ध बादशाह अपने हिस्से की केवल छह भेड़ों के लिए झूठ नहीं बोलेगा ।
And when two such people encounter each other , and their eyes meet , the past and the future become unimportant . There is only that moment , and the incredible certainty that everything under the sun has been written by one hand only .	जब कभी , दो ऐसे व्यक्ति मिलते हैं उनकी नज़रें मिलती हैं , तो अतीत और भविष्य बेमानी हो जाता है । तब रह जाता है केवल वह क्षण और इत्मीनान , एक इतना गहरा यकीन कि दुनिया में जो कुछ भी है , सब कुछ उसी एक के हाथों पहले से लिखा हुआ है ।
It is the hand that evokes love , and creates a twin soul for every person in the world .	वही हाथ प्यार पैदा करता है और दुनिया में हरेक के लिए जुड़वां आत्मा ।
Without such love , one ' s dreams would have no meaning .	ऐसे प्यार के बिना किसी के सपनों का कोई अर्थ नहीं है ।
Maktub , thought the boy .	" मकतूब ! " लड़के ने सोचा ।
The Englishman shook the boy :	अंग्रेज ने लड़के को झकझोरते हुए कहा ,
" Hmm . . . " said the old man , looking at all sides of the book , as if it were some strange object .	" हूं ... " , कहते हुए बूढ़े ने किताब को इस तरह उलटा - पलटा जैसे वह कोई अजीबोगरीब वस्तु हो ।

Zarovnaný paralelní korpus (1658)

Invitatio. **Einleitung.**

M. Veni, Puer! diſce Sapere.

P. Quid hoc eſt, *Supere?*

M. Omnia, quæ *neceſſaria*, rectè *intelligere*, recte *agere*, rectè *eloqui*.

P. Quis me hoc docebit?

M. Ego, cum DEO.

L. Komm her / Knab! lerne Weißheit.

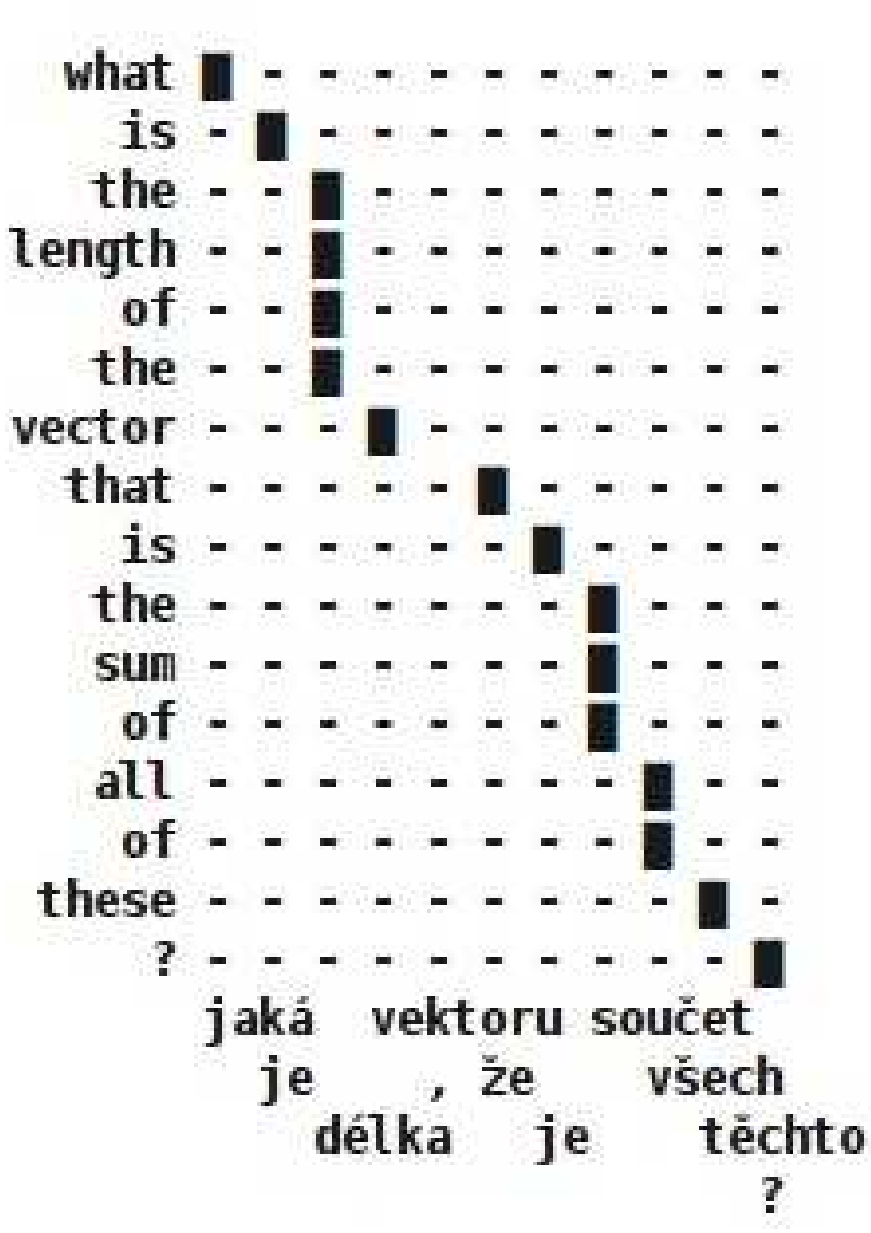
S. Was iſt das / Weißheit?

L. Alles / was nöhtig iſt / recht verſtehen / recht thun / recht ausreden.

S. Wer wird mich das lehren?

L. Ich / mit GOtt.

⇝ Automatické zarovnání po slovech

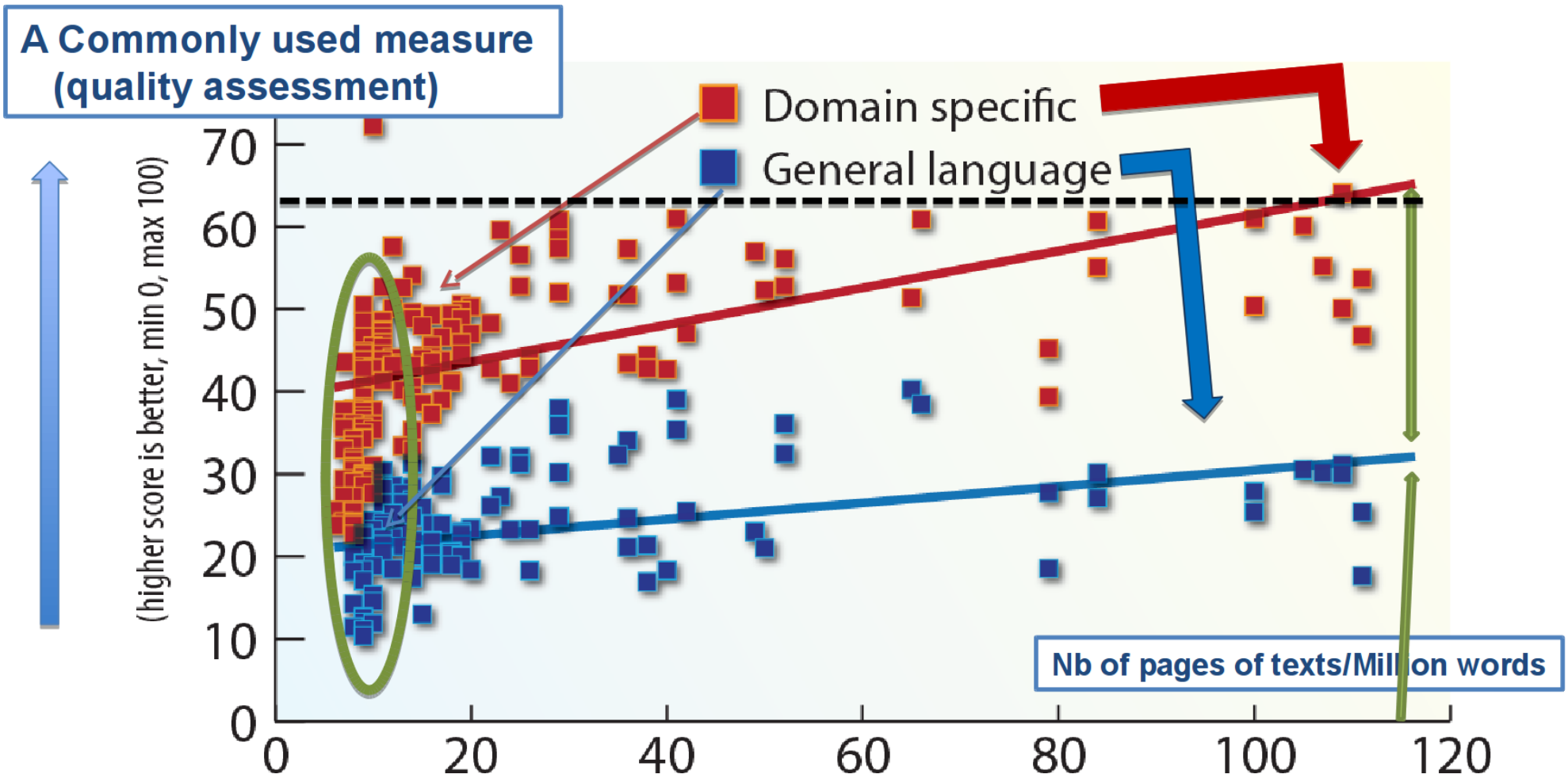


Kolik textů je potřeba?

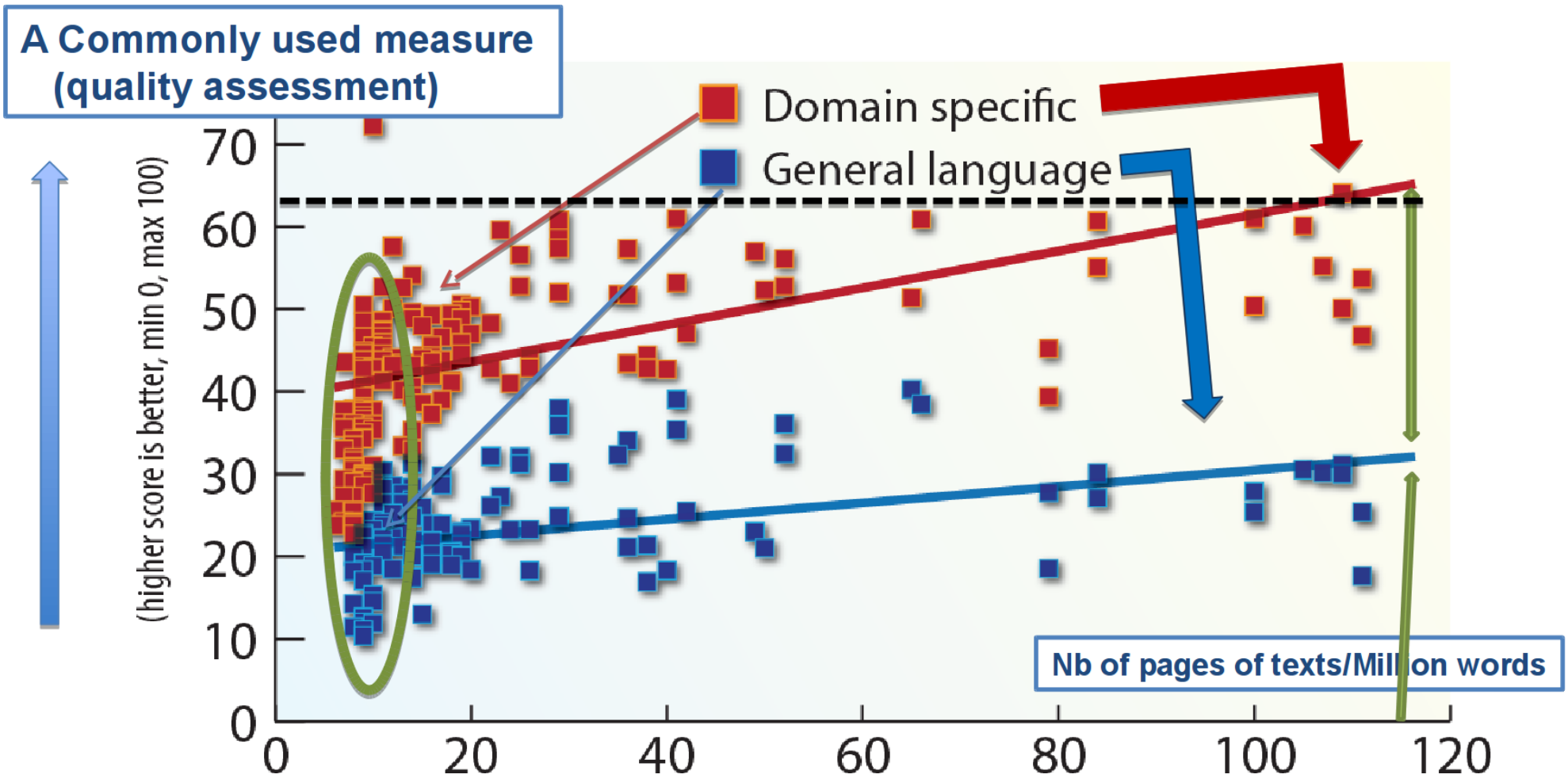
Kolik textů je potřeba?

Víc.

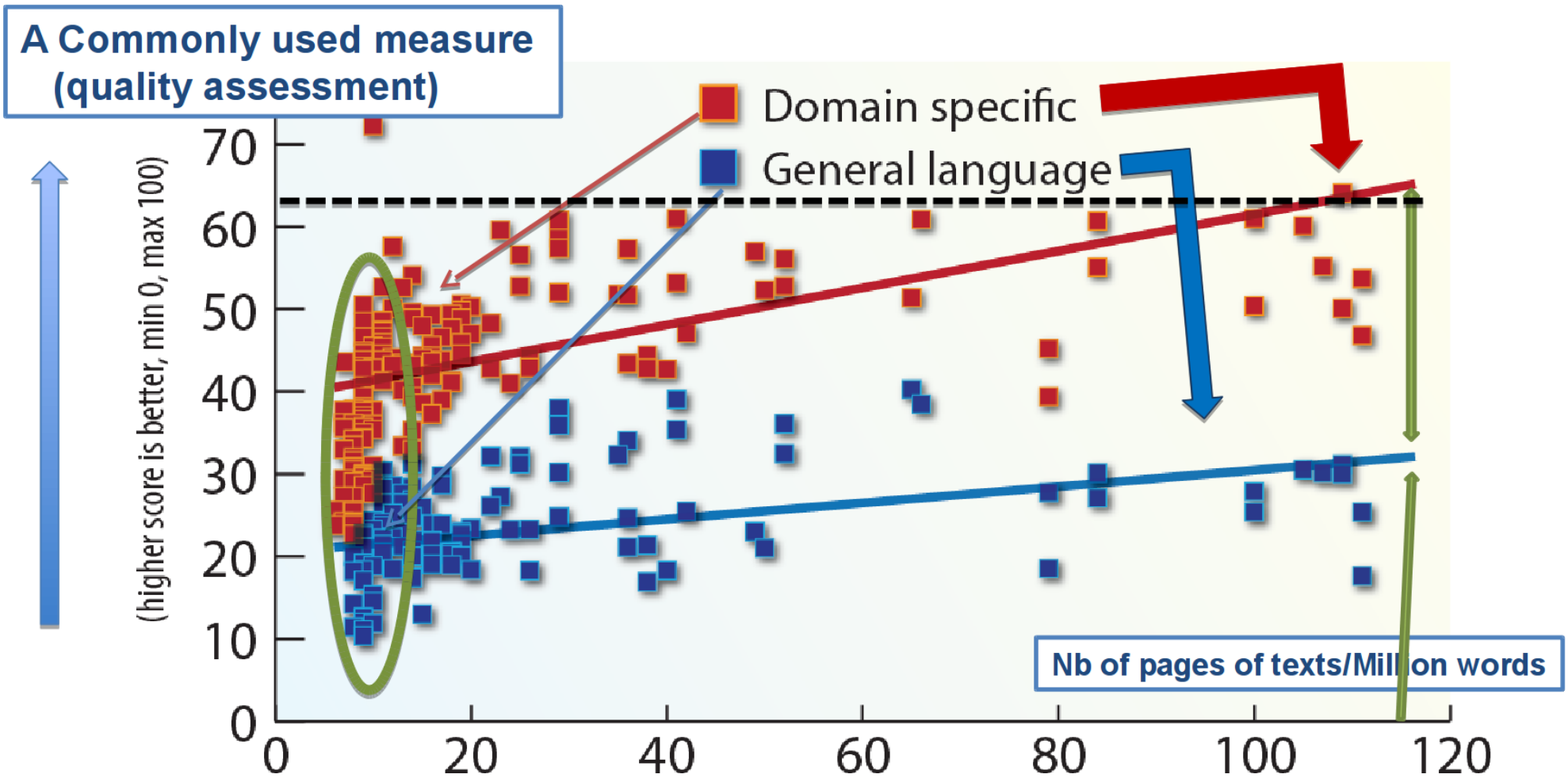
Víc textů ⇒ vyšší kvalita



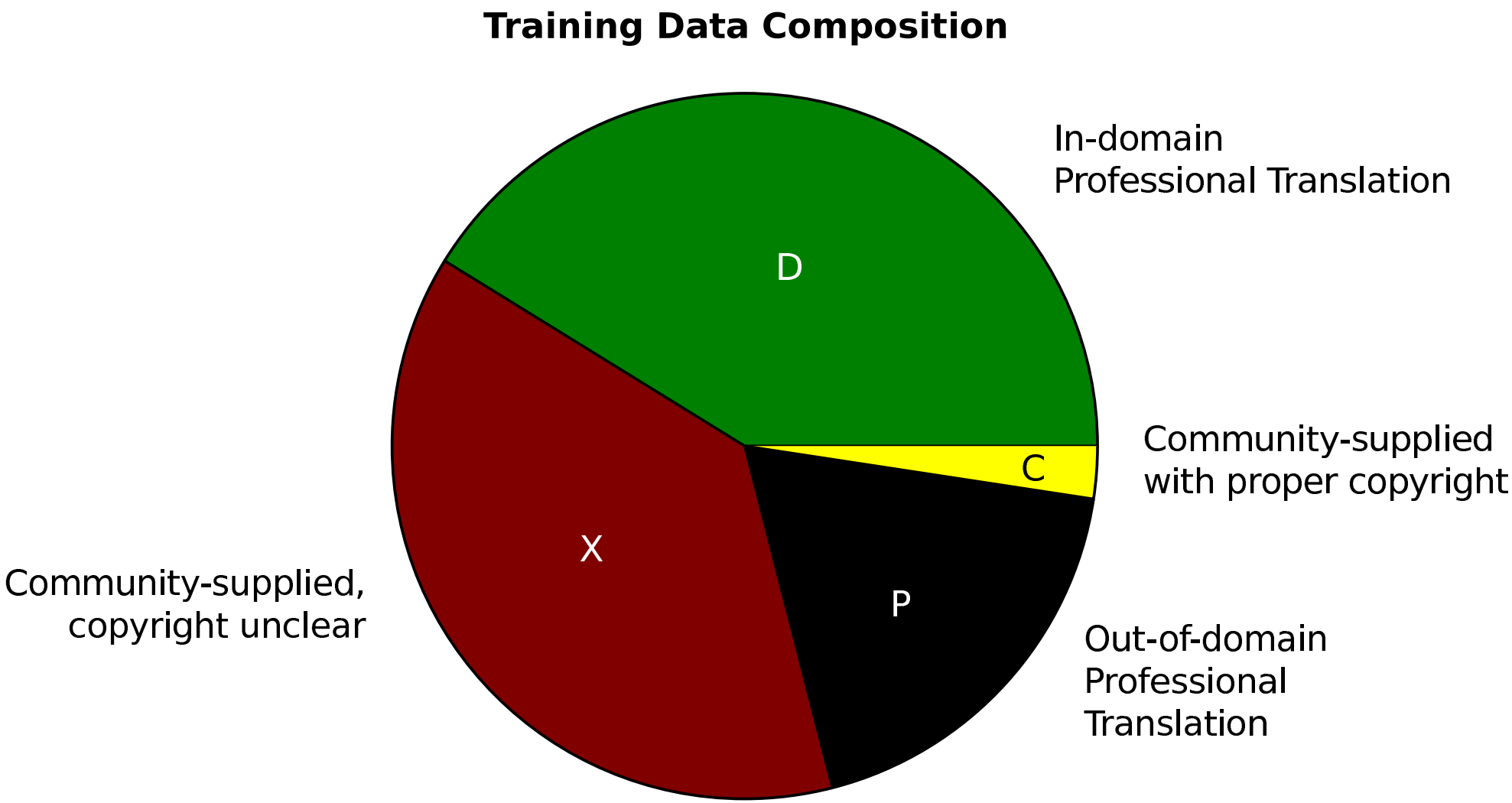
Texty v doméně ⇒ rovnou lepší výstup



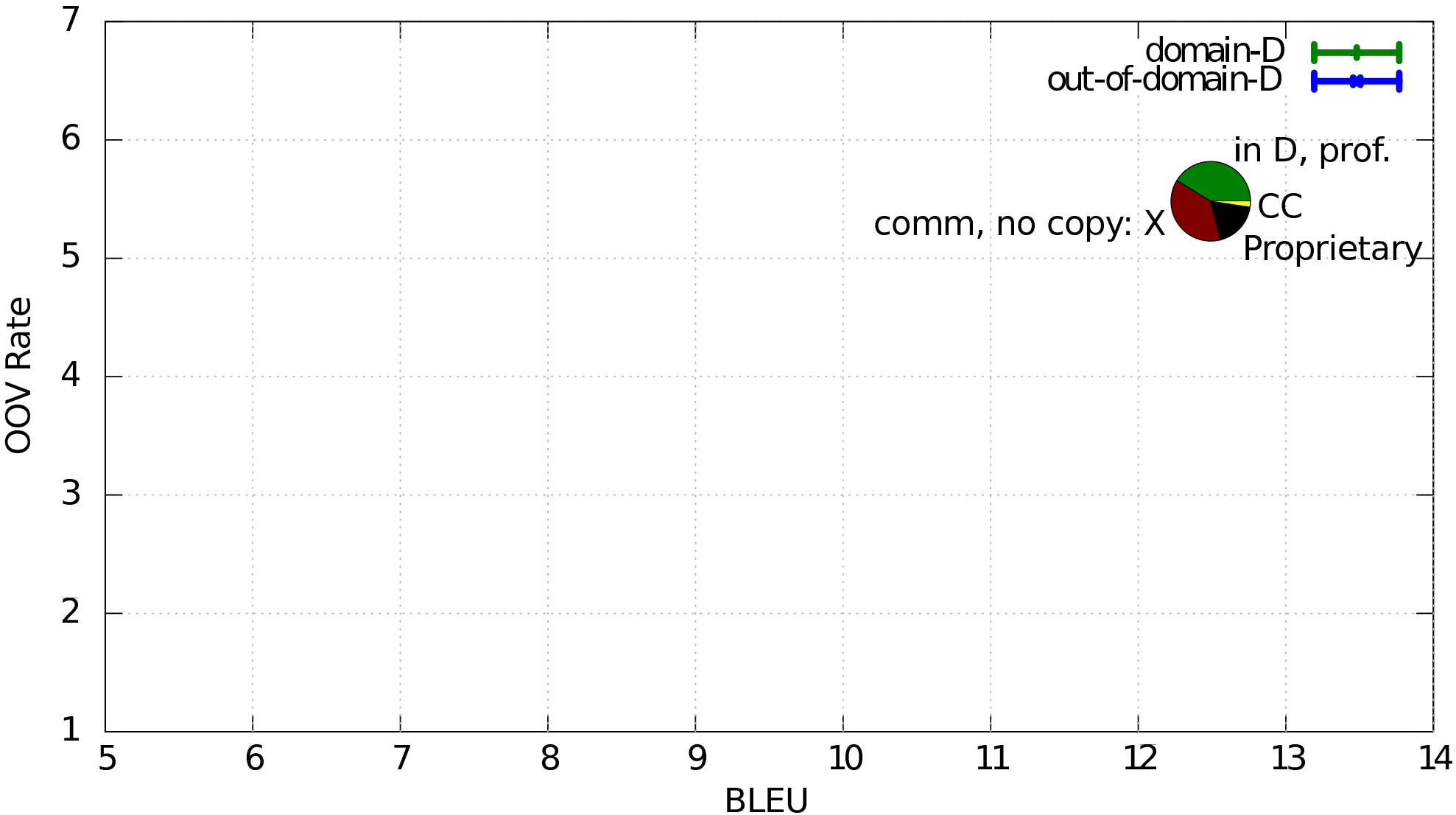
Texty v doméně ⇒ rychlejší zlepšení



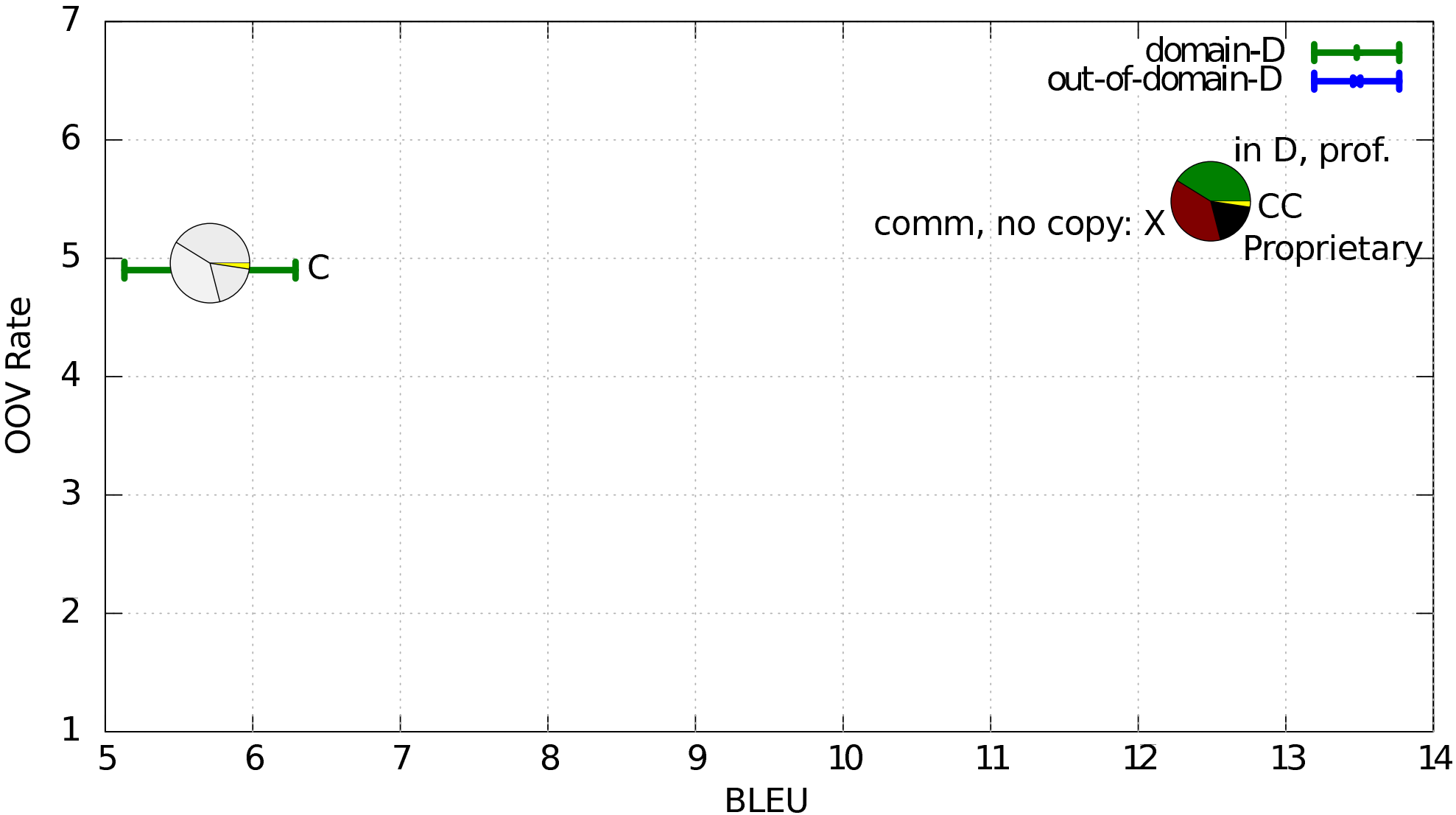
Data pro aj→čj v roce 2008



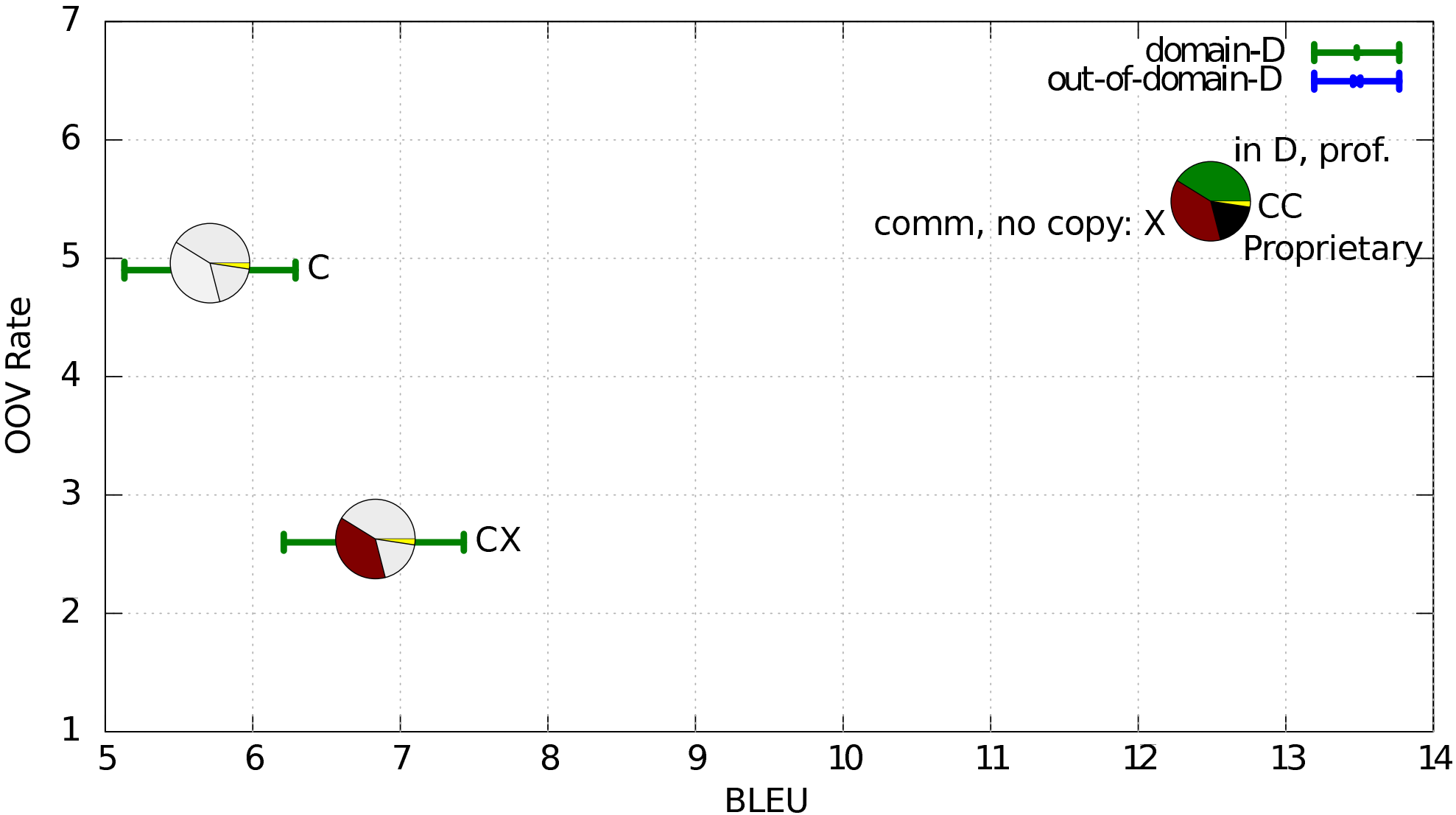
Vliv dat na kvalitu a neznámá slova



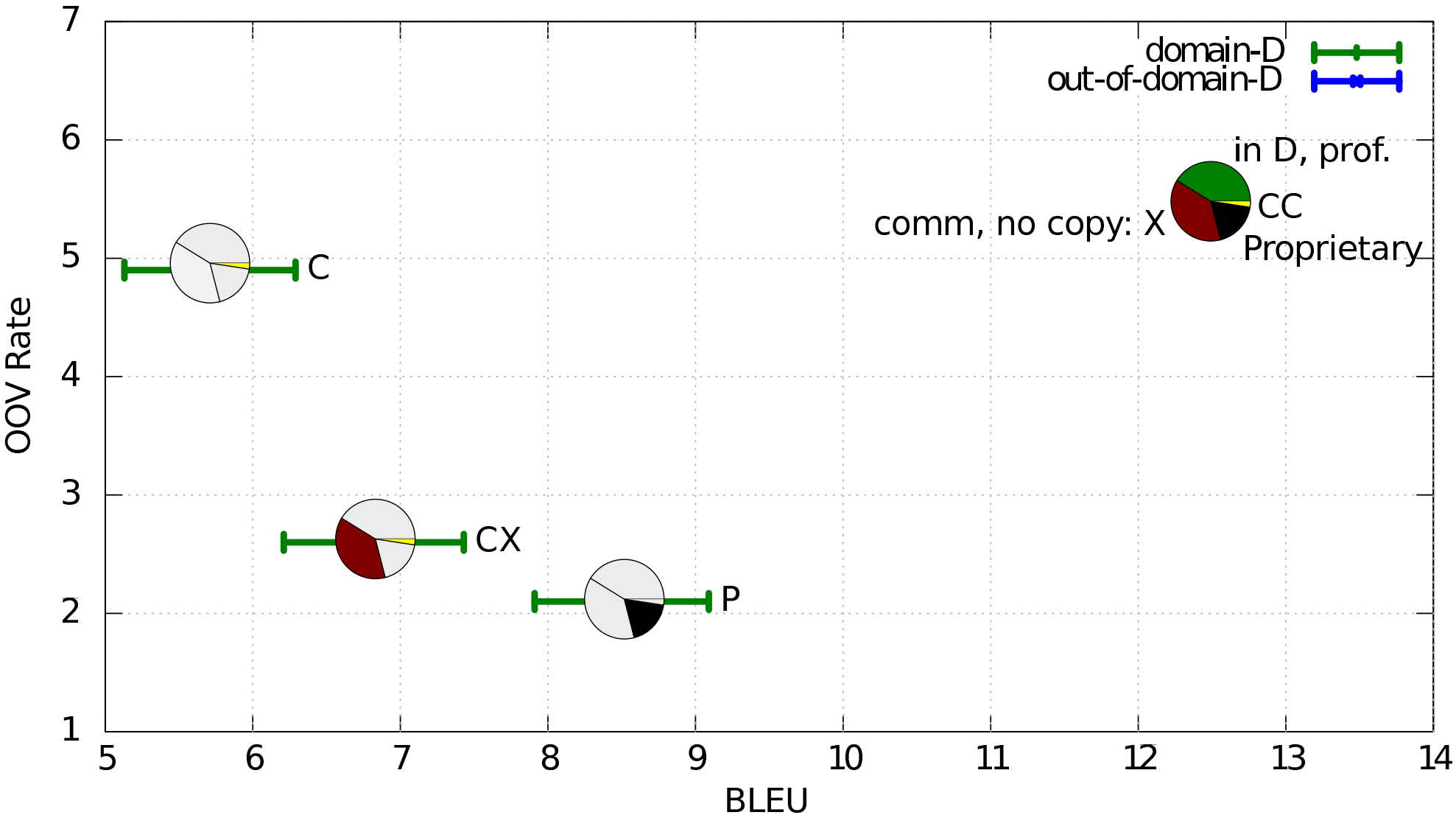
Komunitní data mimo doménu



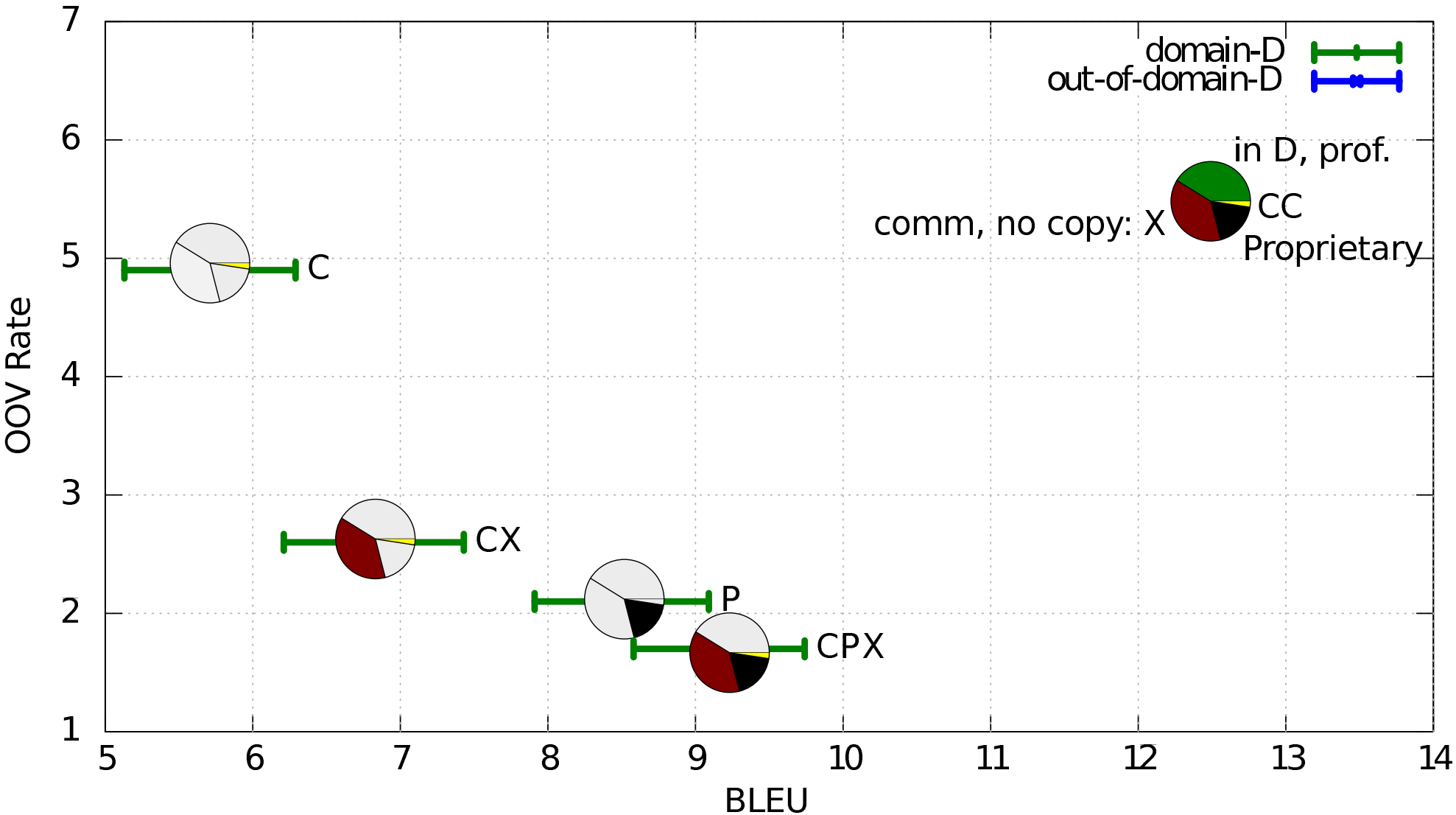
Komunitní data mimo doménu



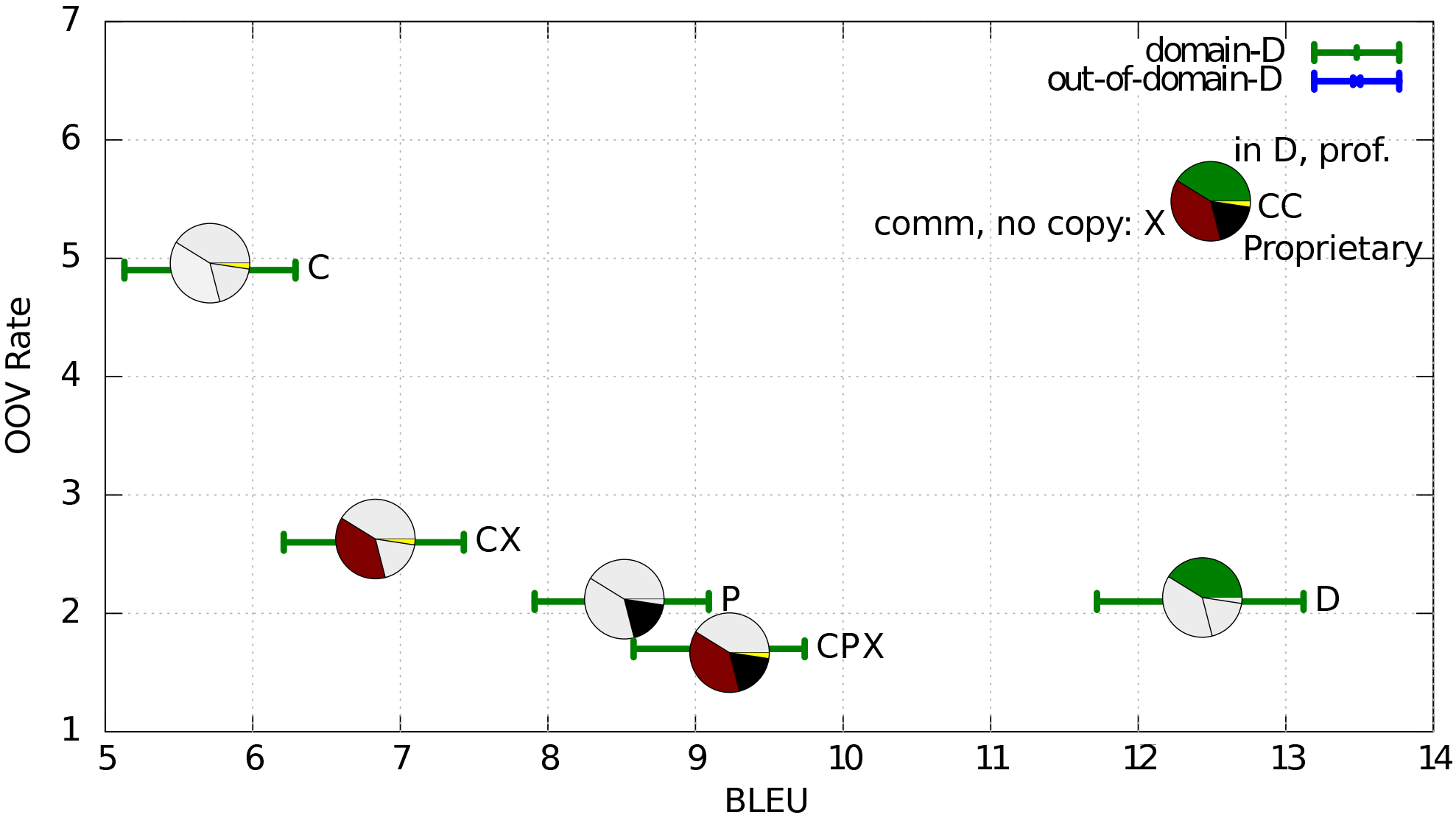
Profesionální překlady mimo doménu



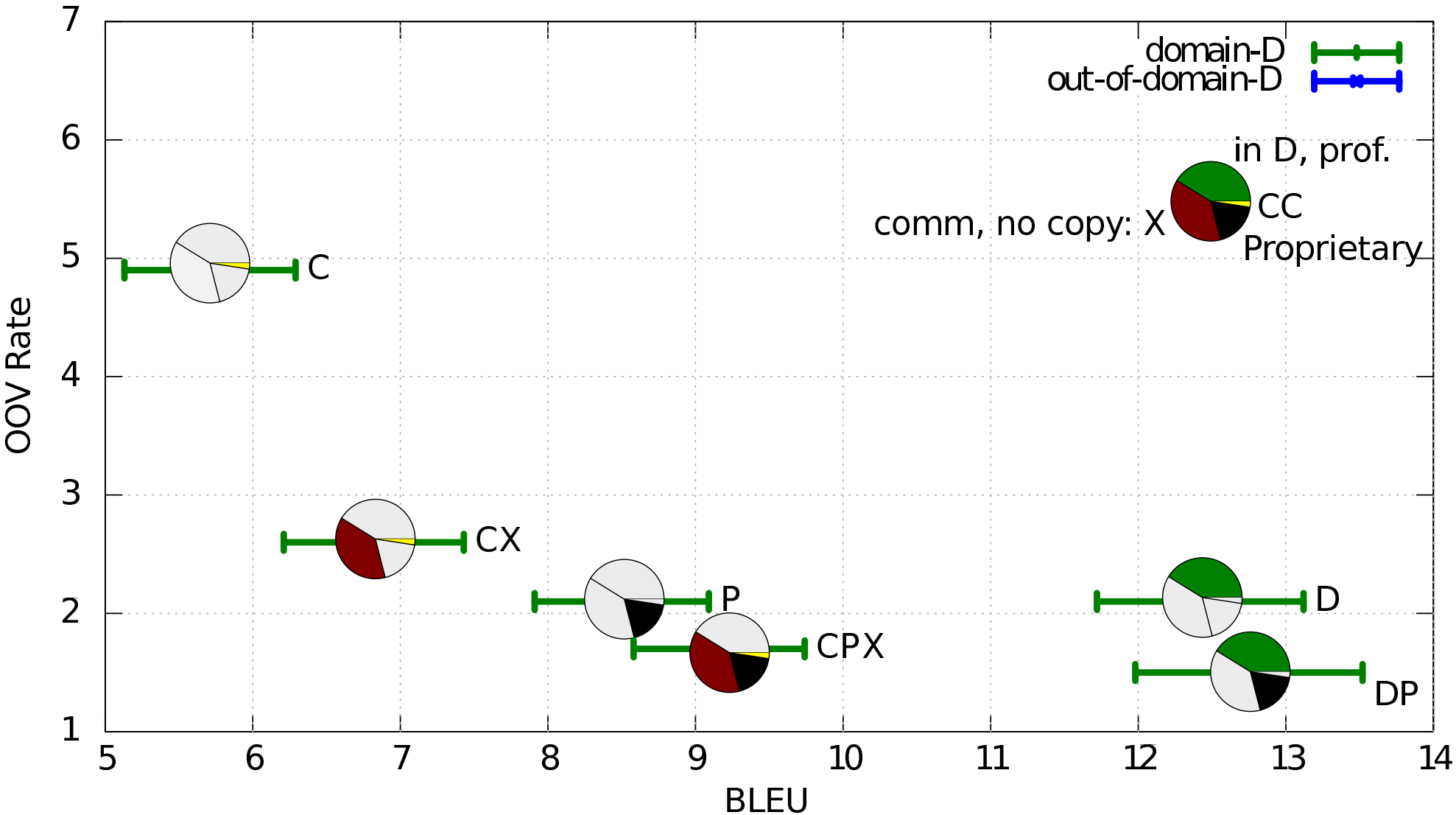
Vše mimo doménu



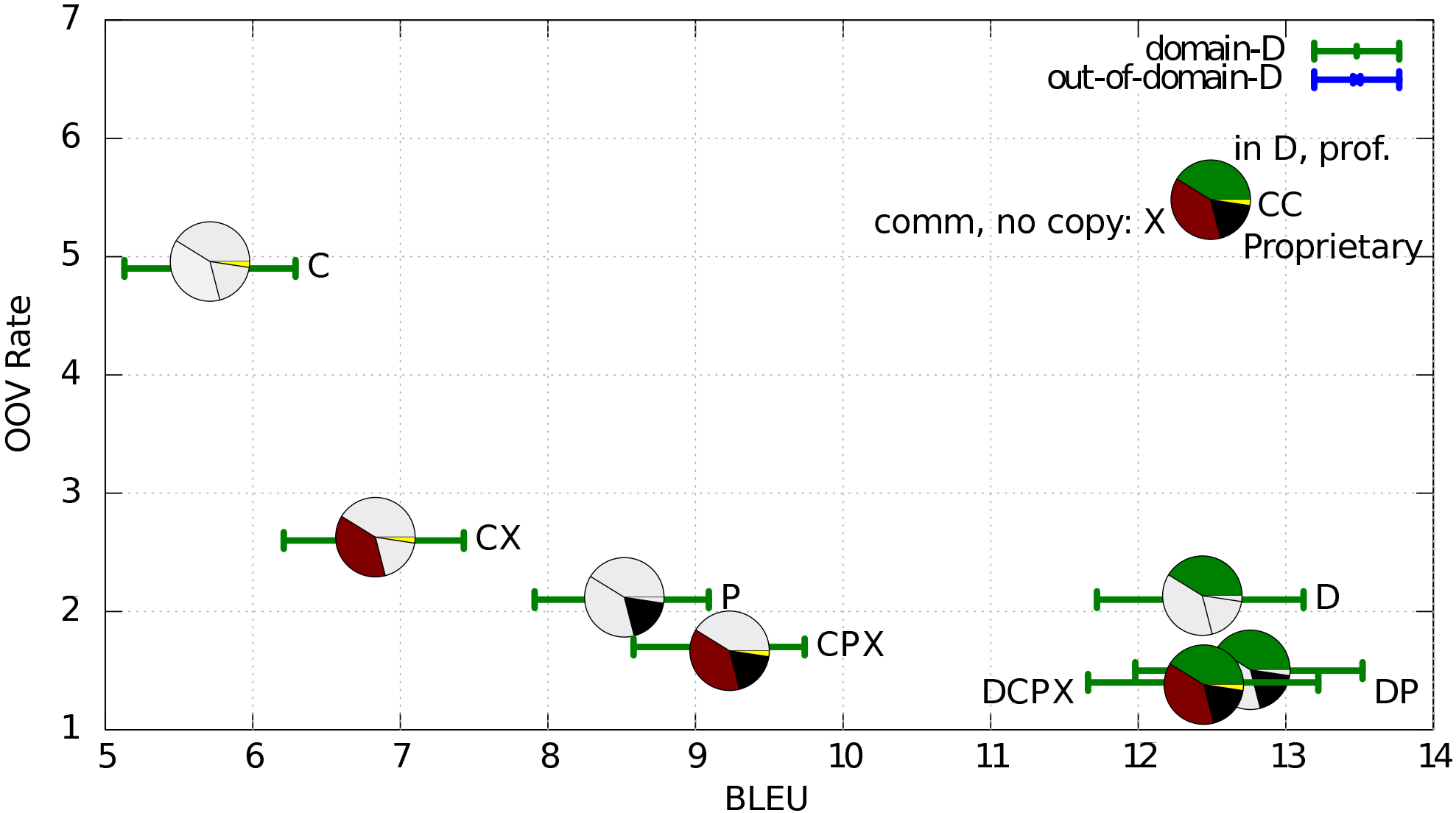
Obdobný objem v doméně: mnohem lepší



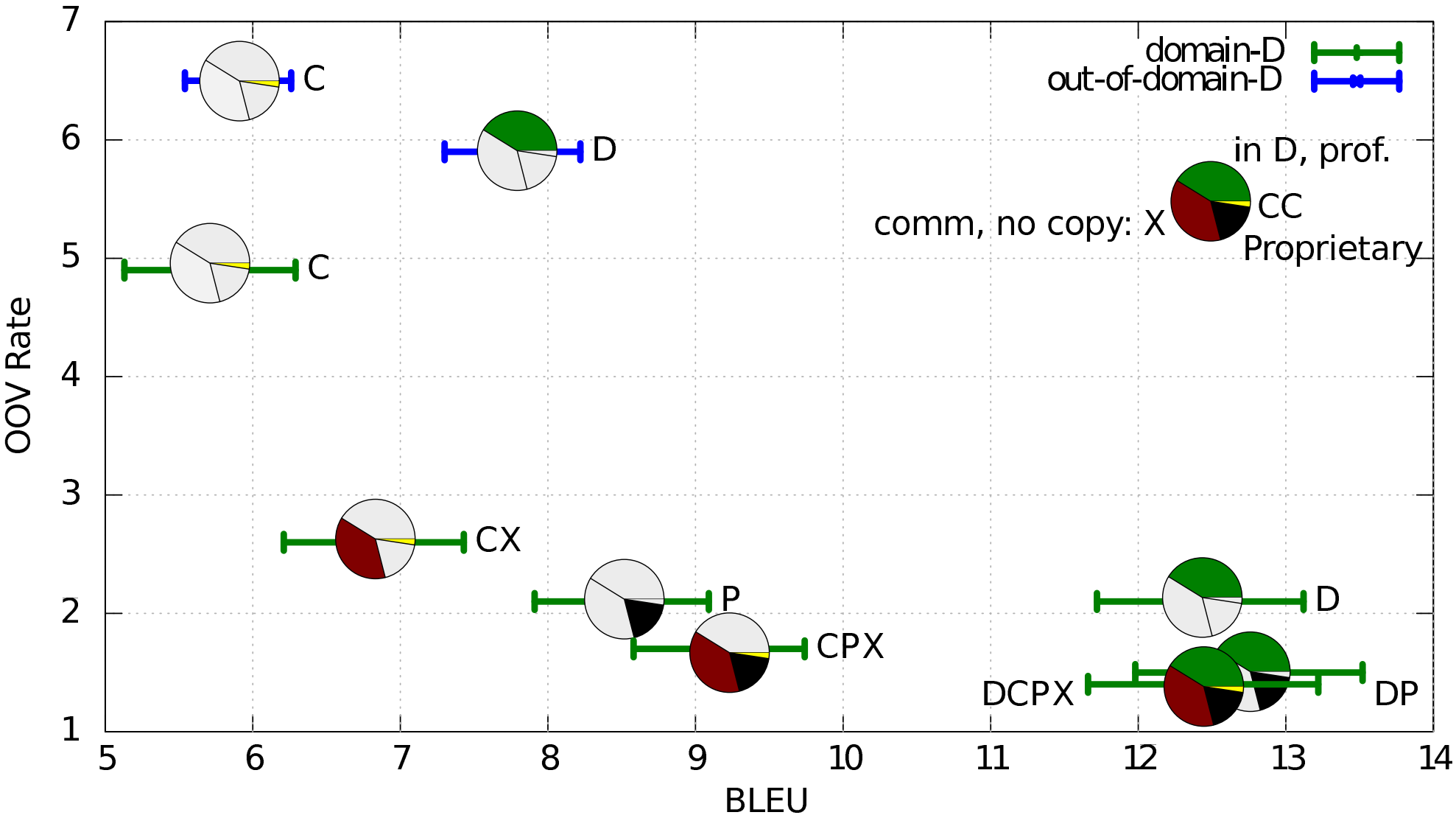
Dodatečná data zlepší pokrytí



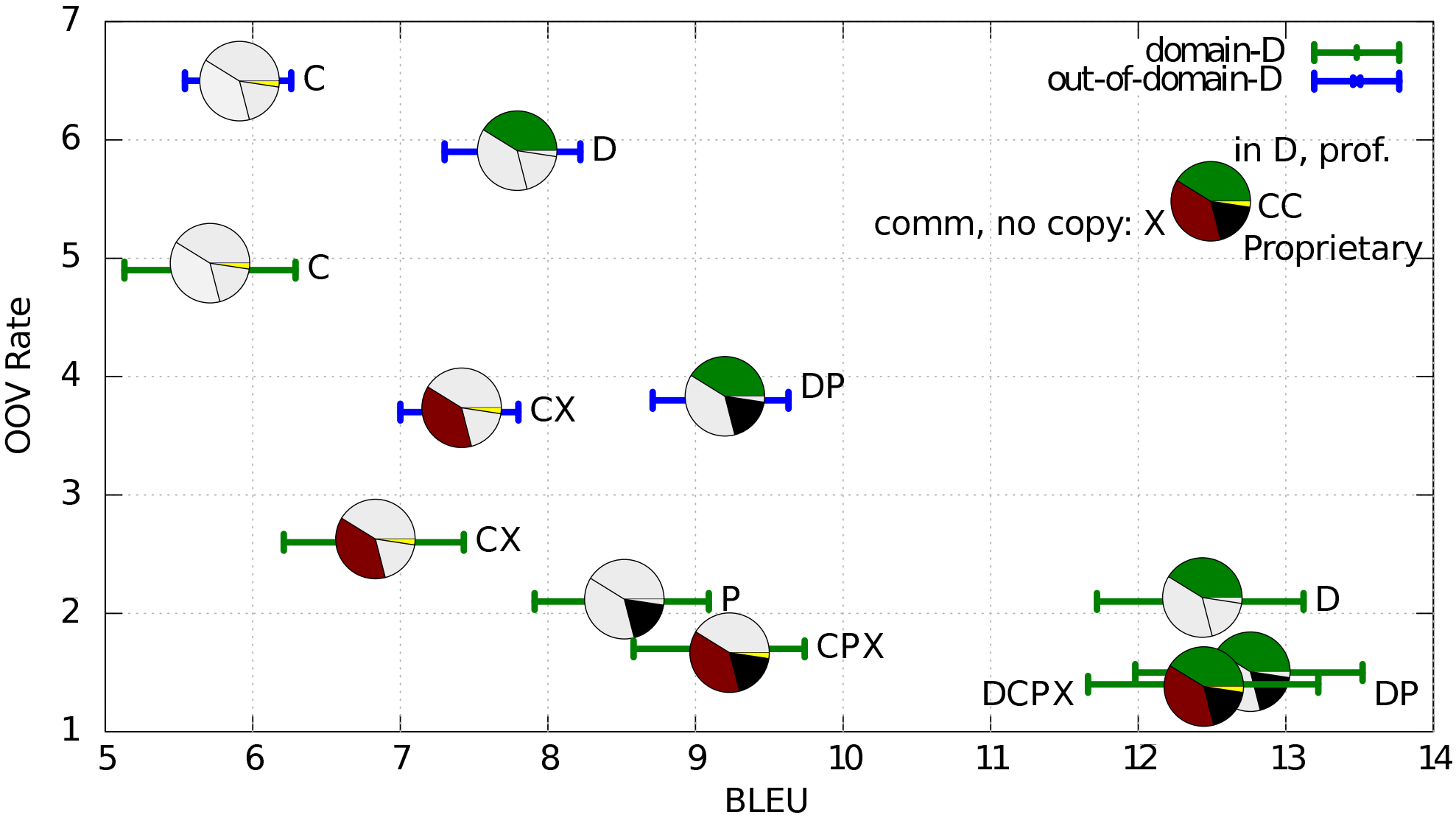
. . . ale mimo doménu mohou snížit kvalitu



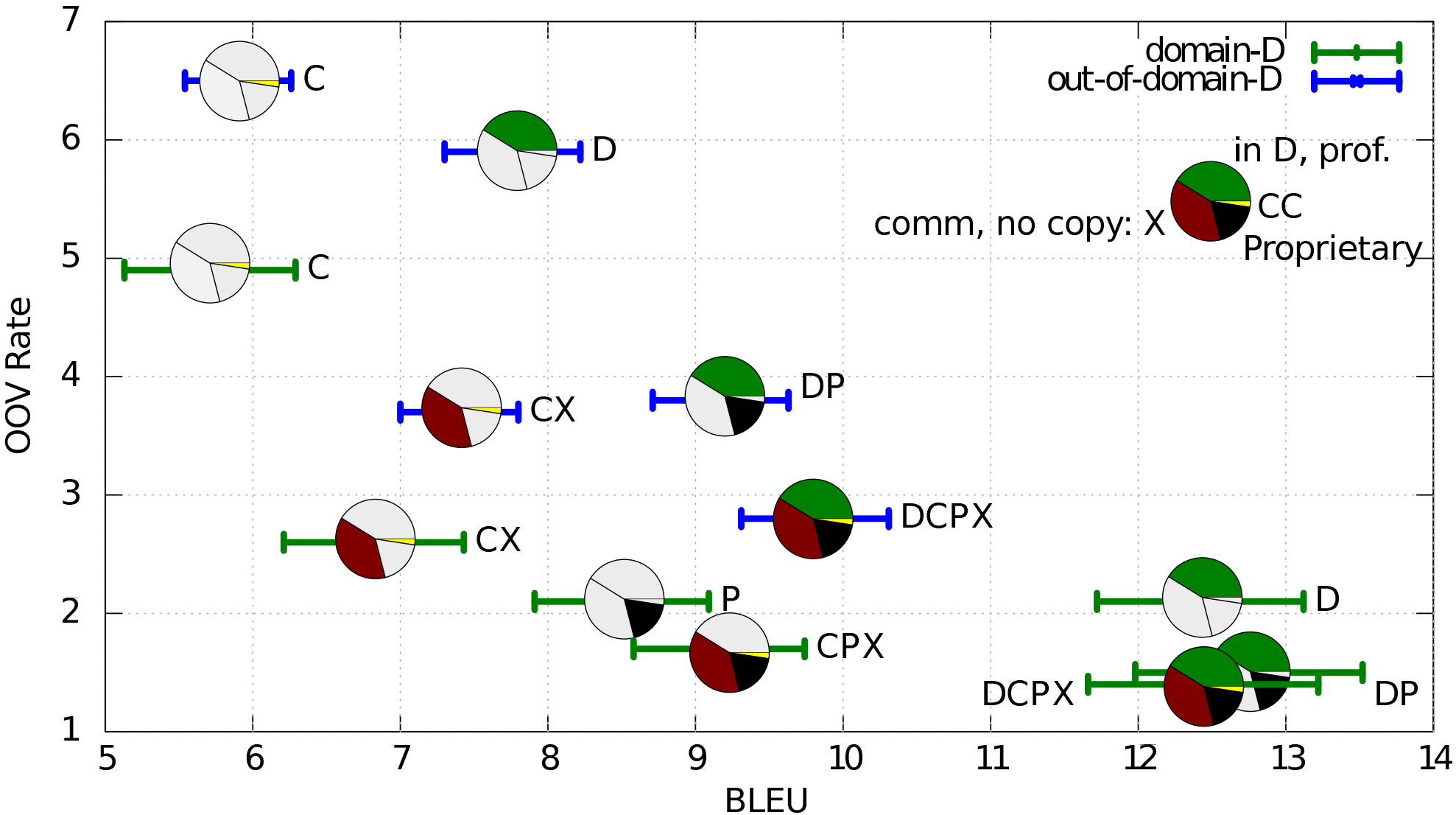
Mimo doménu: kvalita i pokrytí horší



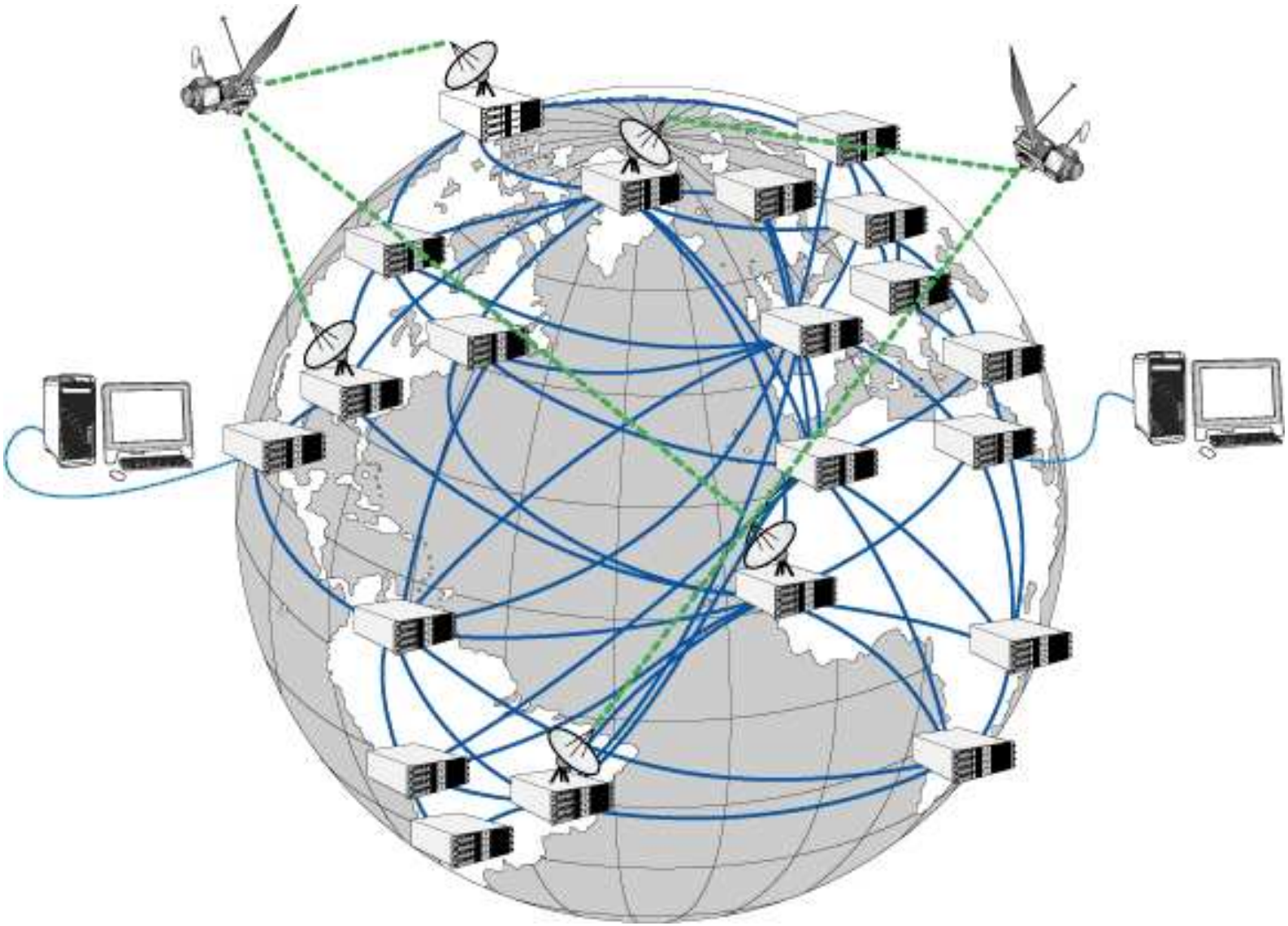
. . . víc dat zlepší pokrytí



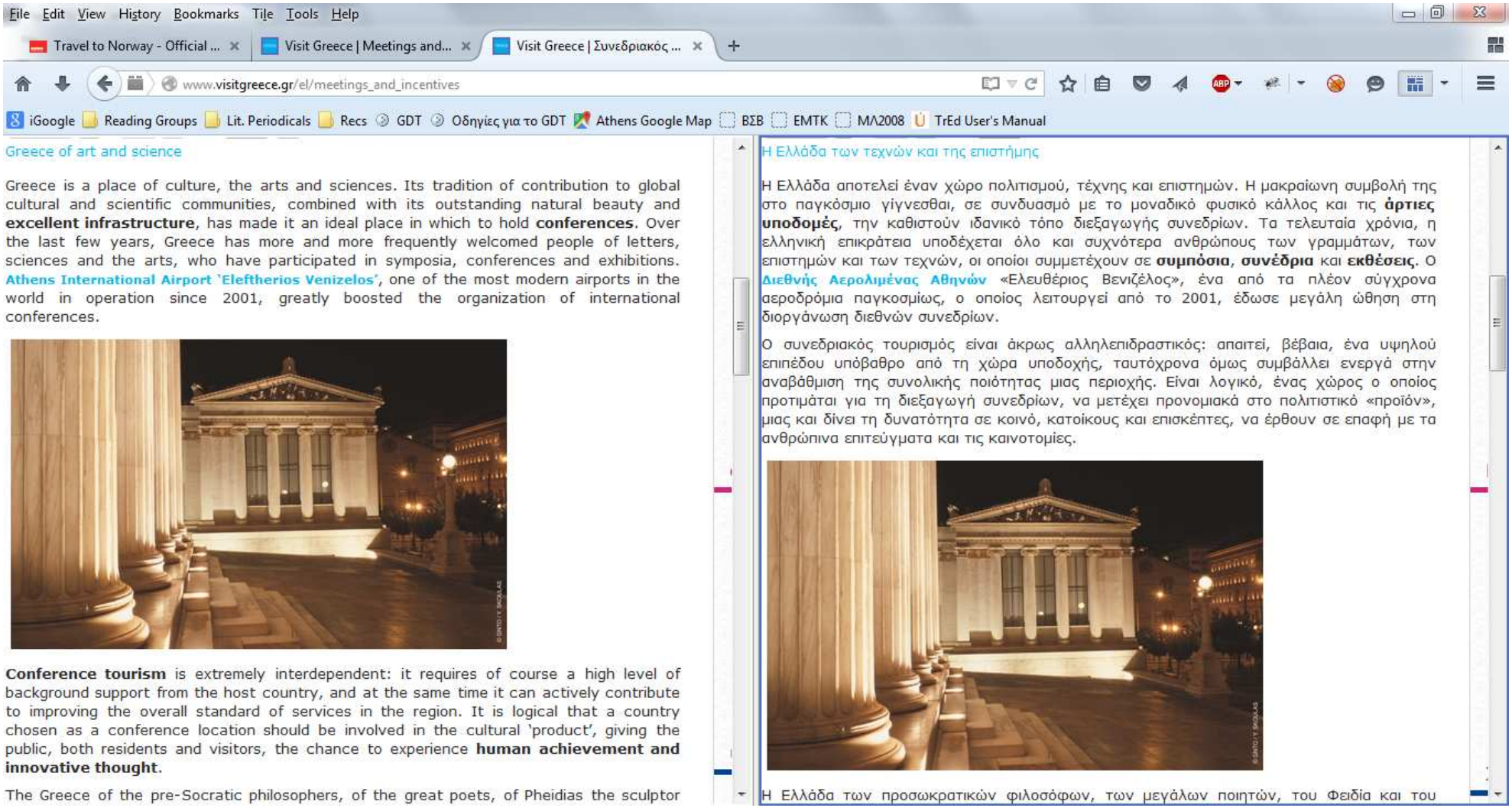
. . . ale na moc kvalitu



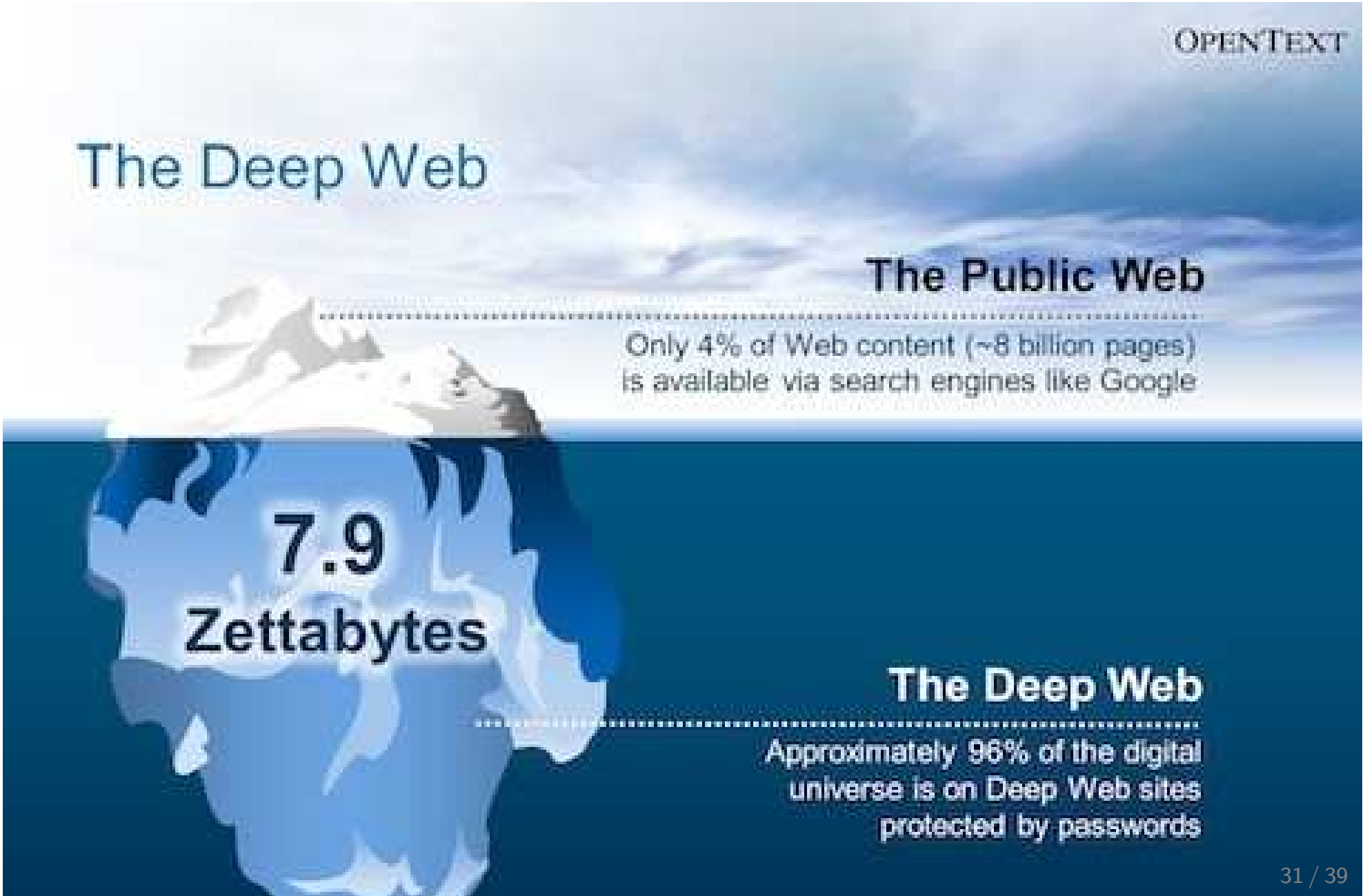
Data raději z webu než z papíru



Paralelní webové stránky



Podstatná část dat nedostupná



Podstatná část dat nedostupná

I dostupné texty nemusí být vytěžitelné

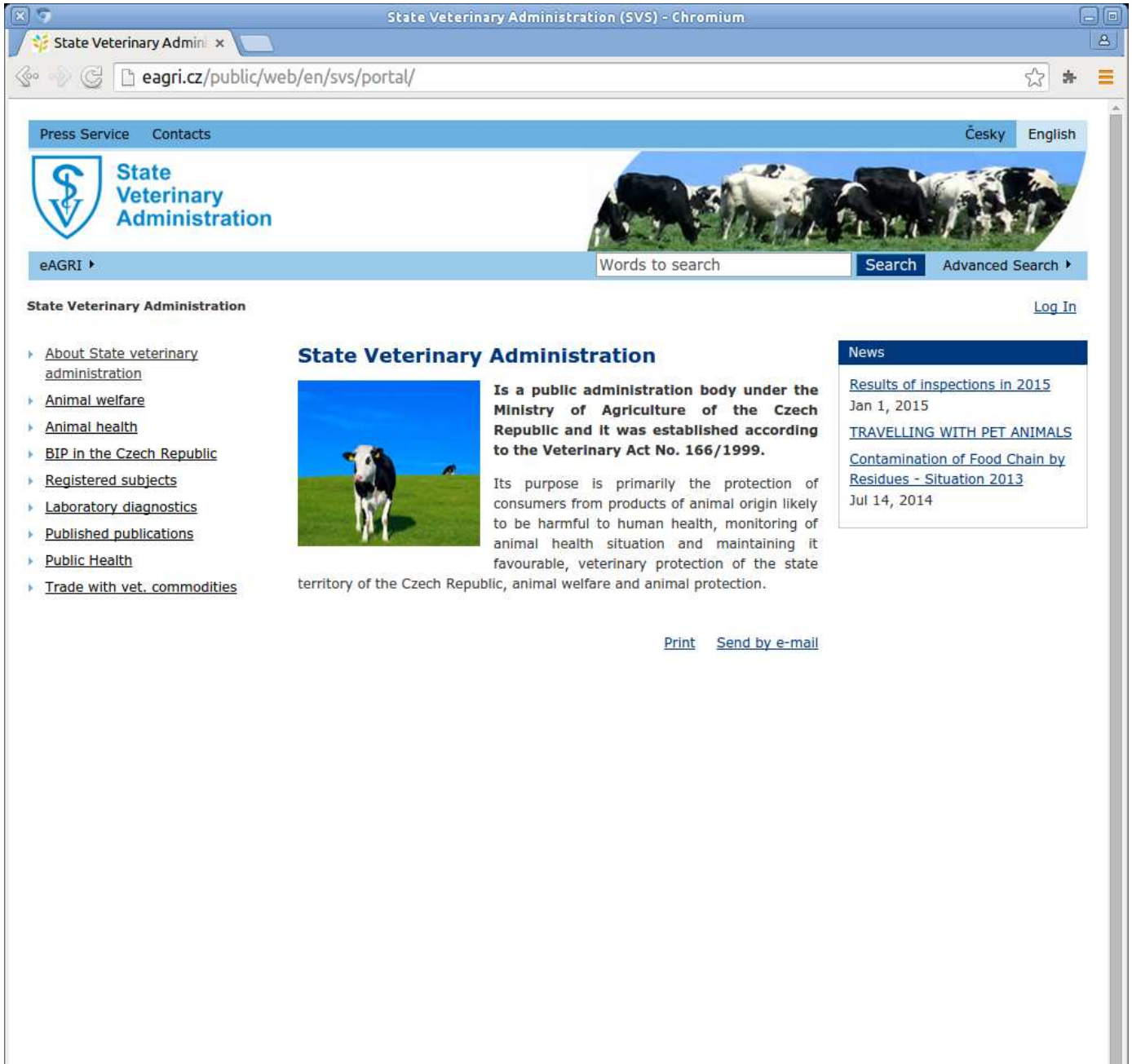
Národní referenční laboratoře SVS ČR

National Reference Laboratories of the SVA CR

Národní referenční laboratoř *National Reference Laboratory*	Sídlo *Address*
I. NRL pro potraviny a krmiva / *NRL for Food and Feed*	
1. NRL pro mléko a mléčné výrobky / *NRL for milk and milk products*	Státní veterinární ústav Praha *State Veterinary Institute Prague* Sídlištní 136/24 165 03 Praha 6 – Lysolaje
2. NRL pro salmonely / *NRL for Salmonellae*	Státní veterinární ústav Praha *State Veterinary Institute Prague* Sídlištní 136/24 165 03 Praha 6 – Lysolaje
3. NRL pro Listeria monocytogenes / *NRL for Listeria monocytogenes*	Státní veterinární ústav Jihlava *State Veterinary Institute in Jihlava* Rantířovská 93 586 05 Jihlava

A mnoho paralelních webů…

. . . je paralelních jen zčásti.



Postup

1. Najít dobré zdroje.
2. Získat surová data.
3. (Identifikovat a pojmenovat domény.)
4. Vyčistit.
   - Odstranit navigační prvky, reklamu, ...
   - Odstranit neparalelní části.
   - Odstranit jiné jazyky.
5. Zarovnat po dokumentech.
6. Zarovnat po větách.
7. Odstranit nekvalitně zarovnané části.
8. Zpřístupnit ostatním.

   $\Rightarrow$ Paralelní korpus CzEng.

`http://ufal.mff.cuni.cz/czeng`

Šlo by to i lépe

- Můžete nám poradit dobré zdroje.
- Můžete nám pomoci s určením domény.
- Můžete nám přímo poskytnout své texty.

Šlo by to i lépe

- Můžete nám poradit dobré zdroje.
- Můžete nám pomoci s určením domény.
- Můžete nám přímo poskytnout své texty.

Shrnutí

YOUR DATA IS NEEDED!

Shrnutí

- Lingvistické analýzy potřebují anotovaná data v doméně.
- MT potřebuje **paralelní texty** v doméně i mimo ni.
- Čím víc, tím líp.
- Čím podobnější vstupům, tím lepší.

- Jazykové zdroje je třeba sdílet.
  - I průmysl data sdílí (např. TAUS).

- Zveřejněné překlady vytěžujeme, netriviální úsilí.
  - Ideálně ne PDF, nejraději HTML s podobným URL.
- Mnohem lepší by byly primární zdroje
  - Vaše texty, překladové paměti.