

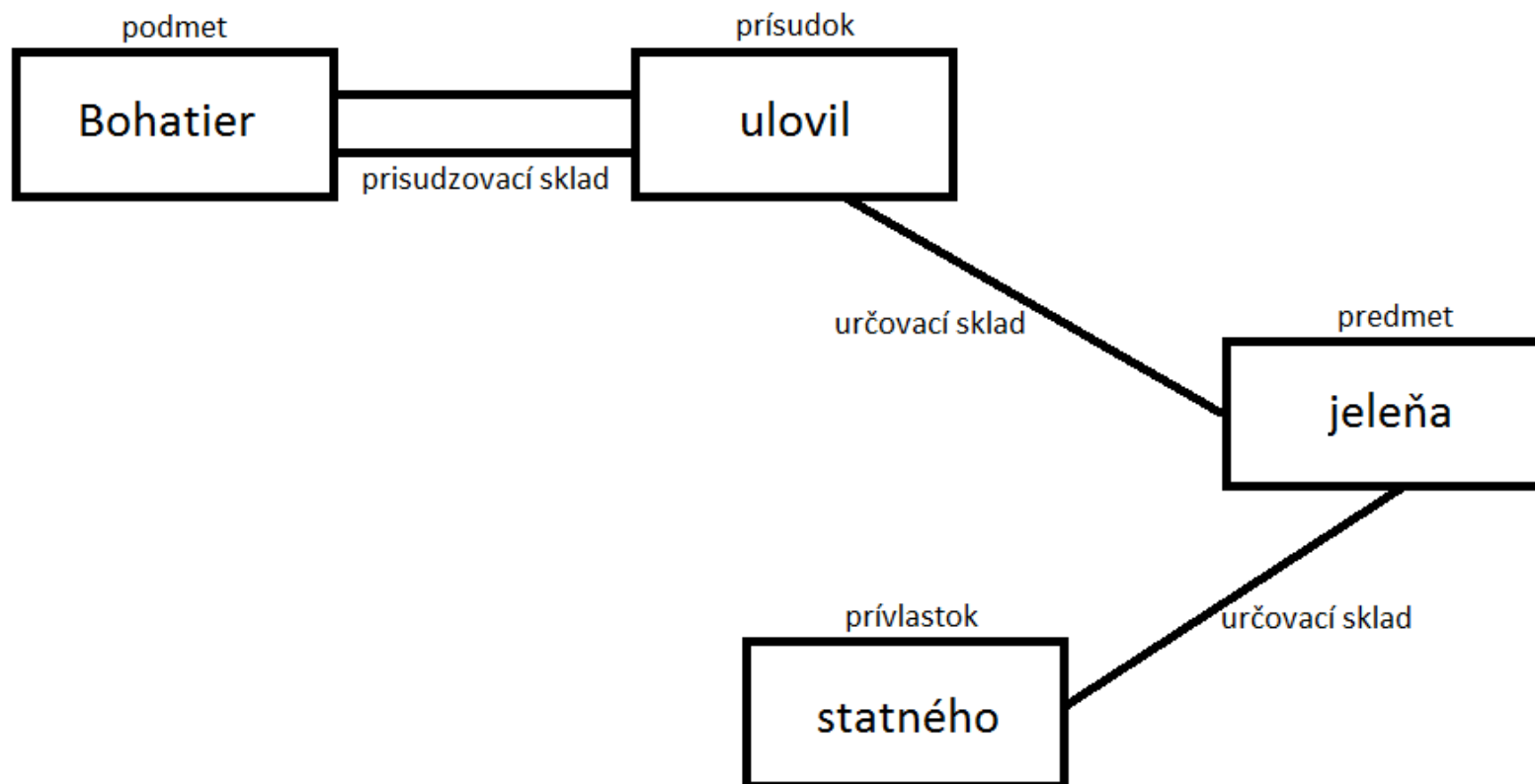
Anotácia rozsiahlych textov za využitia sily davu

Rozmery práce

Spracovanie jazyka – syntaktická analýza

Čerpanie z davu

Bohatier ulovil statného jeleňa



Motivácia (1)

Stav dát pre metódy a techniky spracovania slovenského jazyka je nevyhovujúci.

Využitie čerpania z davu na úlohy NLP

- Dáta pre psycholingvistiku (Munro et al. 2010)
- Preklad textov (Ambati et al. 2010)
- Zjednotenie významu slova, rozlíšenie podobnosti slov, zoradenie udalostí v čase,... (Snow et al. 2008)

Motivácia (2)

Množstvo žiakov na školách, ktorí robia anotácie tak, či tak (209 133 + 76 711).

Konzultácie s učiteľmi – pozitívne prijatie nápadu

- Uľahčenie práce pri distribúcií a kontrole úloh
- Možné zvýšenie motivácie žiakov aplikáciou na pc

Vytvorený dataset sa môže použiť pri ďalších úlohách NLP.

Ciele práce

Vyhodnotiť možnosti využitia sily davu žiakov na získavanie syntaktických anotácií.

Identifikovať granularitu anotácií, ktoré dokážeme získať.

Vyhodnotiť presnosť, kvalitu takto získaných anotácií.

Overiť metódu v softvérovom prototype.

Návrh metody

Metóda na získavanie syntaktických anotácií (1)

Motivácia	známka z domácej úlohy
Kontrola kvality	zhoda vo výstupe, učiteľ
Agregácia dát	postupné budovanie datasetu
Ľudská schopnosť	znalosť syntaktickej analýzy v slovenskom jazyku
Vykonávanie procesu	requester – worker
Kardinalita úloh	many-to-many

QUINN, Alexander J.; BEDERSON, Benjamin B. Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011. p. 1403-1412.

Metóda na získavanie syntaktických anotácií (2)

1. Podkladový korpus
2. Definícia úloh
 1. Typ úlohy (kompletný rozbor, určenie iba vetných členov, ...)
 2. Redundancia počtu riešení
3. Distribúcia úloh davu žiakov
4. Určenie správneho riešenia kombináciou riešení
 1. Nastavenie prahu minimálnej zhody
5. Anotovaný korpus

Metóda na získavanie syntaktických anotácií (3)

Parametre n , e

Berieme väčšinové riešenie

→ Miera zhody $> n$

→ Ostatné elementy $< e$

Budzogáň

Budzogáň Úlohy

admin Odhlásiť

Zoznam úloh

Odovzdať

Veta na rozbor:

Domorodci si delfíny veľmi vážia.

Vetné členy

podmet

prísudok

predmet

prívlastok

príslovkové
určenie



delfíny



Vlastnosti

Vetný člen

predmet

Doplňujúca vlastnosť

Uložiť vlastnosti

Overenie metódy

Pilotný experiment

- Trieda 20 žiakov
- 1 vyučovacia hodina
- 186 anotácií
- Anotácie od učiteľa – zlatý štandard
- 11 viet kde $| \text{anotácie} | > 12$
- DATA-TEST – vety z učebníc slovenského jazyka
- Kvalitatívny experiment

Pilotný experiment

	Správne anotácie
Vetné členy	92,68 %
Vetné sklady	90,00 %
Vety - vetné členy	85,71 %
Vety – kompletne	71,43 %

Experiment 2

180 viet

- Krátke (1 – 5 slov)
- Stredné (6 – 15 slov)
- Dlhé (16 a viac slov)

226 žiakov

2 210 anotácií viet

Použité dáta

DATA-SELECTION – 180 viet

- DATA-SNK – vety zo syntaktického korpusu
- DATA-NEWS – vety z denníka N

Metriky

LAS – Labelled attachment score

UAS – Unlabelled attachment score

Precision

Recall

F-score

Sledované kritériá

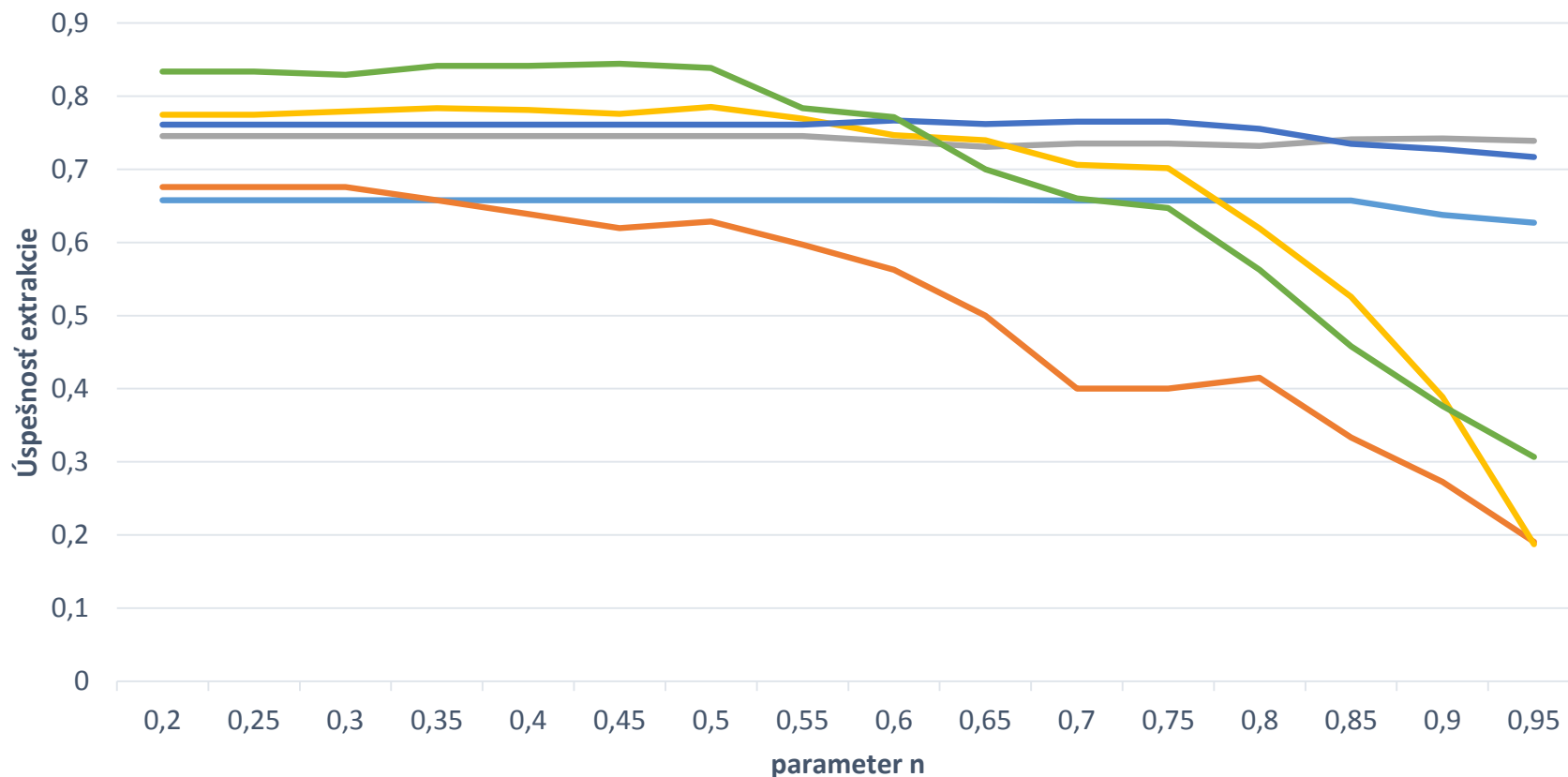
Hodnotenie žiaka

Dĺžka viet

Základná vs. stredná škola

Čas riešenia úlohy (podobný ako na papier 1,50 min)

Dĺžka viet + váhovanie



f-measure (krátke s váhovaním)

f-measure (stredné s váhovaním)

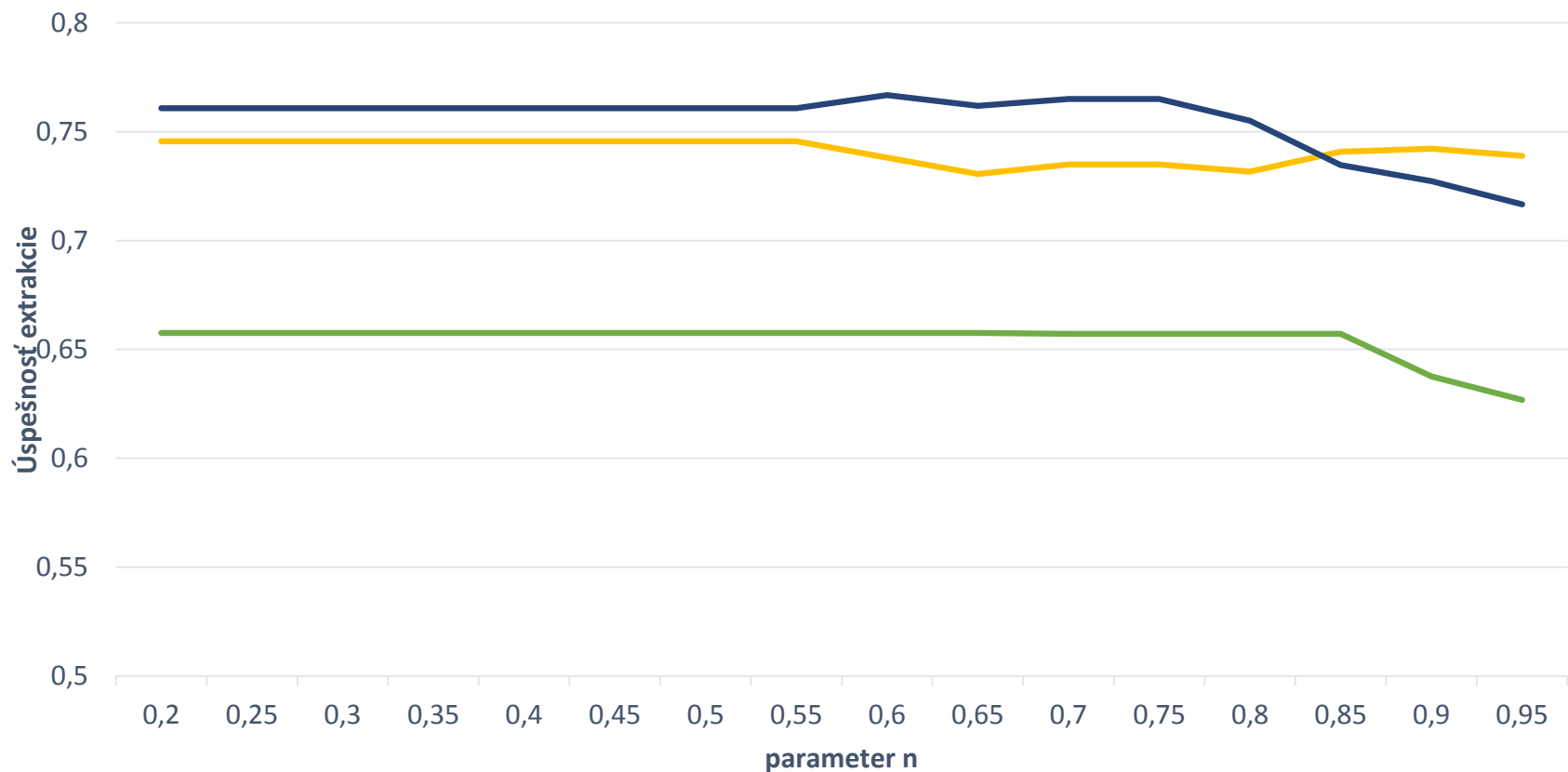
f-measure (dlhé s váhovaním)

f-measure (krátke bez váhovania)

f-measure (stredné bez váhovania)

f-measure (dlhé bez váhovania)

Dĺžka viet

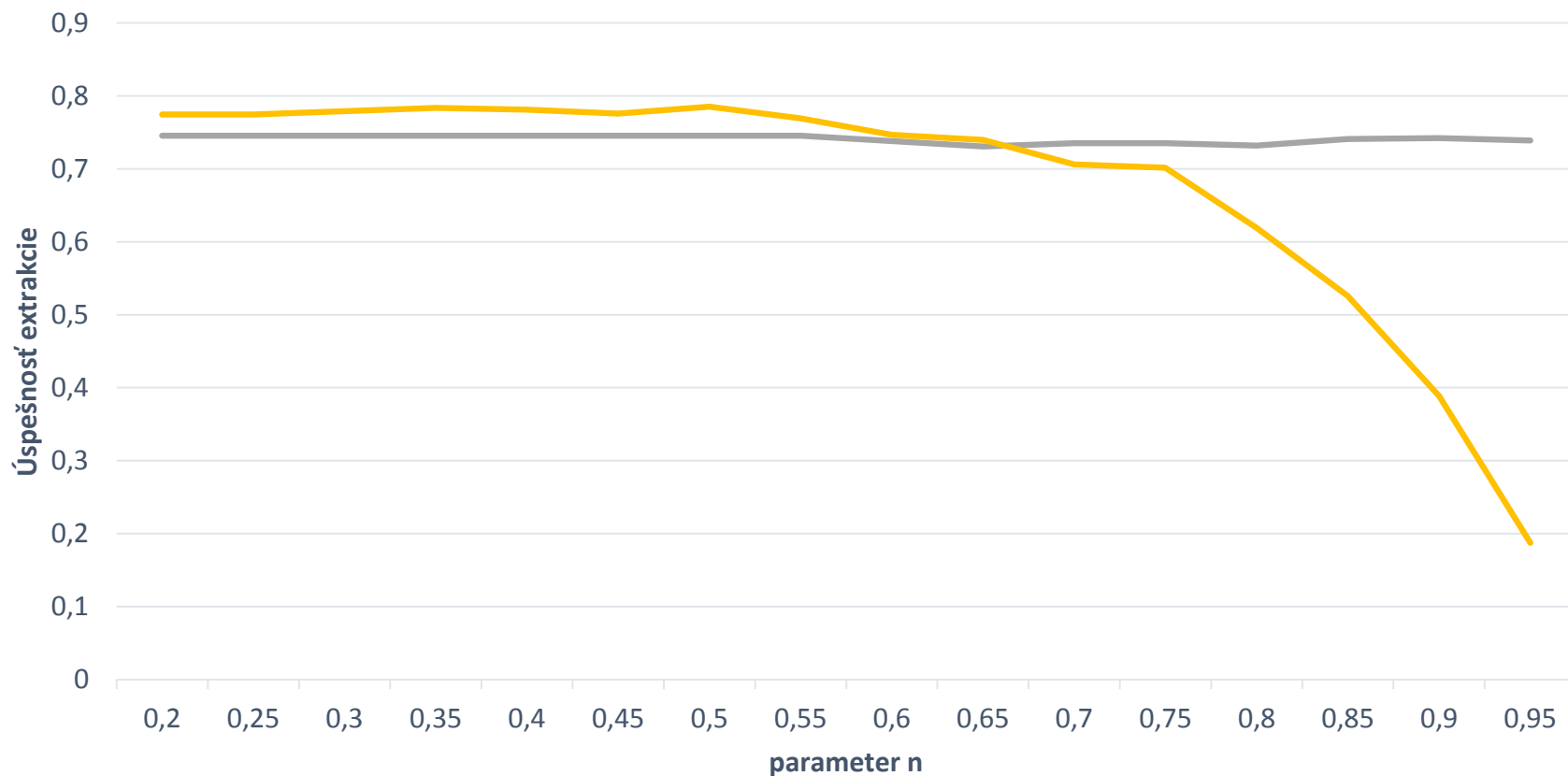


— f-measure (krátke s váhovaním)

— f-measure (stredné s váhovaním)

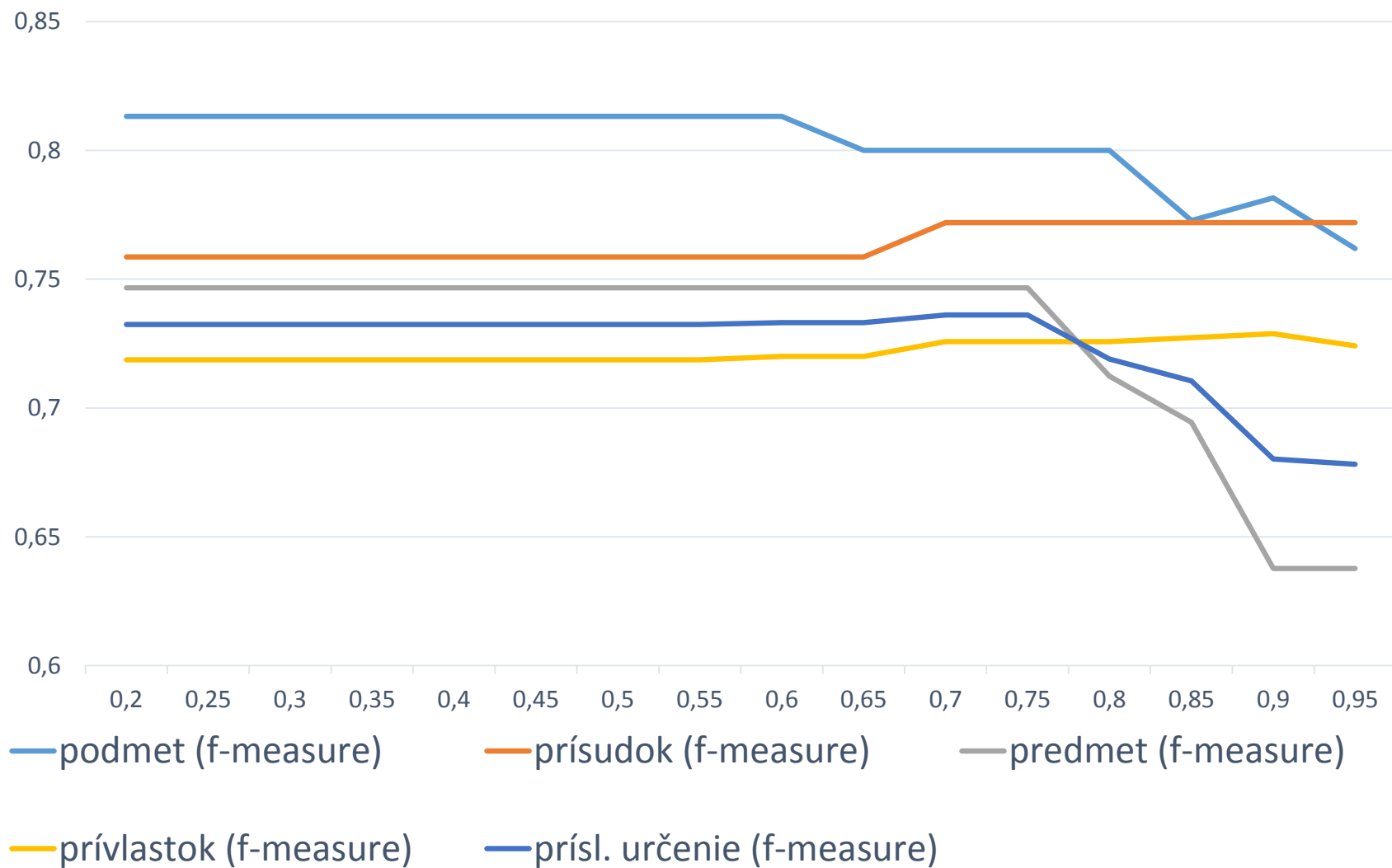
— f-measure (dlhé s váhovaním)

Váhovanie



— f-measure (stredné s váhovaním) — f-measure (stredné bez váhovania)

Vetné členy



Základná vs. stredná škola

n = 0,75	Precision	Recall	F-measure
Základná škola	0,80	0,56	0,66
Stredná škola	0,69	0,35	0,46

Vetné sklady

Parameter n	UAS – krátke vety (%)	UAS – stredné vety (%)	UAS – dlhé vety (%)	LAS – krátke vety (%)	LAS – stredné vety (%)	LAS – dlhé vety (%)
0,2	88,89	53,33	40,16	66,67	28,89	36,07
0,25	88,89	35,56	27,87	66,67	22,22	23,77
0,3	77,78	28,89	20,49	66,67	20,00	16,39
0,35	66,67	22,22	14,75	55,56	13,33	12,30
0,4	66,67	8,89	11,47	44,44	6,67	8,20
0,45	66,67	6,67	3,28	44,44	6,67	2,46
0,5	66,67	6,67	2,46	44,44	6,67	1,64

Zhrnutie

Návrh a overenie metódy využívajúcej dav žiakov na získavanie syntaktických anotácií.

Pozitívna spätná väzba od učiteľov a žiakov.

Syntaktické anotácie s menšou kvalitou a granularitou ako od lingvistov.

Najlepšie výsledky pre podmet a prísudok, vetné sklady v krátkych vetách.

IIT.SRC 2015

Budzogáň

Webová aplikácia

→ Ruby on Rails

→ Bootstrap

→ PostgreSQL

Prostredie na tvorbu syntaktických anotácií.

Prostredie pre učiteľa – definícia úlohy

Záver

Návrh a overenie metódy na tvorbu syntaktických anotácií.

Syntaktické anotácie s menšou kvalitou a granularitou ako od lingvistov.

Možné použitie na podporu niektorých úloh NLP.

Pozitívna spätná väzba od učiteľov a žiakov.

IIT.SRC 2015

Otázky (1)

V práci sa diplomant nevyjadruje k tomu ako navrhnuté riešenie závisí od konkrétneho jazyka. Je riešenie použiteľné aj pre iný ako slovenský jazyk?

Navrhnuté riešenie je nezávislé od jazyka, avšak samotná realizácia už je závislá od slovenčiny.

Otázky (2)

V diskusii v časti 7.1.2 (s.39) diplomant uvádza „V pilotnom experimente sme zistili, že našou metódou dokážeme získavať korektné anotácie za použitia relatívne malého davu (20 žiakov). Je to naozaj tak? Aké sú ohraničenia vykonaného pilotného experimentu?

Dokázali sme úspešne určiť vybrané vetné členy.

Ohraničenia

- Malá množina viet
- Malá vzorka žiakov
- Malý priestor na experimentovanie s redundanciou riešení.

Otázky (3)

Diskutujte možnosti zlepšenia metódy, aj v kontexte veľkosti davu a redundancie riešených úloh.

Väčší dav môže priniest zlepšenie výsledkov – viac dát, väčšia možnosť selekcie výsledkov.