MASARYKOVA UNIVERZITA

FAKULTA INFORMATIKY

Vylepšenie hry X-plain

BAKALÁRSKA PRÁCA

Monika Kostolná

Brno, 2013

Prehlásenie

Prehlasujem, že táto práca je mojim pôvodným autorským dielom, ktoré som vypracovala samostatne. Všetky zdroje, pramene a literatúru, ktoré som pri vypracovaní používala alebo z nich čerpala, v práci riadne citujem s uvedením úplného odkazu na príslušný zdroj.

Monika Kostolná

Vedúca práce: RNDr. Zuzana Nevěřilová.

Poďakovanie

Za pomoc a hodnotné rady pri vytváraní tejto práce ďakujem vedúcej práce RNDr. Zuzane Nevěřilovej. Poďakovanie patrí aj členom laboratória spracovania prirodzeného jazyka Masarykovej univerzity za ich ústretovosť pri zabezpečení potrieb tejto práce. V neposlednom rade sa chcem poďakovať aj rodine a priateľom, ktorí mi boli oporou počas celého štúdia.

Zhrnutie

Práca sa v prvej časti venuje spôsobu získavania vedeckých dát pomocou konceptu Games with a Purpose a bližšie popisuje charakteristické vlastnosti týchto dát. Po stručnom náčrte hry X-plain rozoberá jednotlivé vylepšenia tejto hry ako prepojenie s rozhraním Facebook, vytvorenie slovenskej verzie a iné rozšírenia vedúce k väčšej atraktivite. Hra je pre používateľov spomínanej sociálnej siete Facebook dostupná na internetovej adrese: http://nlp.fi.muni.cz/projekty/x-plain_enhanced/game/ a pre hru mimo nej na adrese: http://nlp.fi.muni.cz/projekty/x-plain_enhanced/.

Kľúčové slová

X-plain, Sketch Engine, Facebook, Common sense facts, korpus, Games with a purpose,

Obsah

1 Úvod ... 3
1.1 Motivácia ... 3
1.2 Cieľ práce ... 4
1.3 Obsah kapitol ... 4
2 GWAP's ... 6
2.1 Common sense facts ... 6
2.2 Vlastnosti dát ... 7
2.3 Gamifikácia ... 8
3 Nástroje NLPC ... 10
3.1 Morfologický analyzátor *majka* ... 10
3.2 Značkované korpusy ... 12
3.3 Sketch Engine ... 14
3.3.1 Concordance ... 14
3.3.2 Word Sketch ... 15
4 Dostupné programovacie prostredie Facebook API ... 17
4.1 Facebook Developers ... 17
5 Hra X-plain – pôvodná verzia ... 18
5.1 Bot ... 18
5.2 Použité nástroje ... 18
5.3 Cieľ hry ... 19
5.4 Preklad relácií *„z Gramatiky do sémantiky"* ... 19
5.5 Súvisiace projekty ... 20
5.5.1 Vedecké práce ... 20
5.5.2 Vedecko-herné koncepty ... 20
5.6 Splnenie vlastností GWAP's ... 21
6 Rozšírenie hry X-plain ... 22
6.1 Terminológia a značenie ... 22
6.2 Objektový návrh hry ... 22
6.2.1 Trieda *Game* ... 22
6.2.2 Trieda *Round* ... 23
6.2.3 Trieda *Player* ... 24
6.2.4 Triedy *ComputerPlayer* a *HumanPlayer* ... 25
6.3 Slovenská verzia ... 25
6.3.1 Pridanie ďalšieho jazyka ... 26
6.4 Inovácie databázy ... 26

6.5 Využitie Facebook API 28
6.5.1 Spracovanie osobných údajov 28
6.5.2 Hra mimo Facebook 29
6.6 Bezpečnosť aplikácie 29
6.7 Použitie herných elementov Gamifikácie 30
6.8 Implementácia prezentácie dát z hry 30
6.8.1 Štatistiky a tabuľky výsledkov 30
6.8.2 Grafické prvky 31
7 Návrhy do budúcnosti 32
8 Záver 33

1 Úvod

1.1 Motivácia

Dôležitou zložkou každého projektu je kooperatívna činnosť. Spolupráca, či už medzi viacerými ľuďmi, alebo v podobe automatizácie medzi človekom a počítačom, je kľúčom k vyriešeniu mnohých problémov. Aj samotných užívateľov, ktorí dielo uvedú do života, môžeme za istých okolností považovať za spoluautorov. V súvislosti s tým vystupuje v posledných rokoch do popredia koncept tzv. *Games with a Purpose (GWAP's)* [1], hier, ktoré za pomoci svojich užívateľov zbierajú rozsiahle dáta, určené na vedecké účely. Takéto hry má zmysel vytvoriť predovšetkým ako pomocný nástroj k projektom, ktorých realizácia je časovo nákladná a vyžaduje si veľké množstvo spracovaných informácií. Príkladmi hier zbierajúcich masívne množstvo dát je napríklad EyeWire, kde užívatelia mapujú neurónové bunky očnej sietnice [2] alebo hra Verbosity zameraná na kumuláciu textových dát spolu so vzťahmi medzi nimi [3].

Druhá spomínaná sa zameriava na zber tzv. *common-sense facts (CSF)*, slovných spojení a výrazov, ktoré sú považované za „bežným človekom poznané". Tieto dáta nachádzajú veľké uplatnenie v oblasti umelej inteligencie, kde sa podieľajú na tvorbe inteligentných systémov, ontológie (definície pojmov a vzťahov medzi nimi) alebo v lingvistických centrách určených na tvorbu vyhľadávačov a nástrojov na dopĺňanie slov.

V kooperácii s členmi Laboratória spracovania prirodzeného jazyka Masarykovej univerzity (*Natural Language Processing Centre NLPC*) prichádzam s projektom X-plain, hrou na webe, ktorá je typickým príkladom GWAP's. V hre za pomoci nástrojov NLPC vytváram automatizovaného hráča, ktorý simuluje buď napovedanie, alebo hádanie slova popisovaného užívateľom. Dôležitý je tu preklad relácií z korpusového manažéra *Sketch Engine* do podoby, ktorá by bola použiteľná v hre pri získavaní tipov a indícií od počítača. Najdôležitejšou súčasťou hry je databáza, obsahujúca vyextrahované CSF z jednotlivých ťahov hráčov, ktoré sú vo forme trojíc

hádané slovo – relácia – tip prístupné ako vedeckým centrám, tak aj užívateľom hry.

1.2 Cieľ práce

Cieľom mojej práce je vylepšenie existujúcej českej hry X-plain [4]. Snažím sa urobiť ju atraktívnejšou pre užívateľov, aby zaujala širšie herné osadenstvo a tak zväčšila databázu CSF.

Hru navrhujem a vytváram odznova, za účelom sprehľadnenia kódu a použitia novších a efektívnejších metód pre jej funkčnosť. Ďalej umiestňujem hru na sociálnu sieť Facebook, s využitím dostupných prostriedkov, ktoré táto sieť ponúka na jej propagáciu. Z dôvodu minimálnych rozdielov medzi češtinou a slovenčinou hrávali doteraz hru aj slovenskí užívatelia, ktorí však priniesli do hry značné chybné dáta, a preto vytváram k hre aj slovenský modul, zbierajúci slovenské CSF. V neposlednom rade je mojou úlohou pridať do hry aj štatistiky a výsledkové tabuľky zvyšujúce atraktivitu hry a tiež zlepšiť zabezpečenie aplikácie.

1.3 Obsah kapitol

Druhá kapitola približuje koncept GWAP's, cieľové dáta, ktoré je z tohto druhu hier možné vygenerovať, a bližšie načrtáva ich požadované rysy. Taktiež popisuje stávajúci fenomén gamifikácie a jeho elementy, z ktorých niektoré aplikujeme aj do našej hry.

Tretia kapitola oboznamuje s nástrojmi NLPC, ktoré boli použité buď ako pomocné prostriedky pri vytváraní tejto aplikácie, alebo ako základ funkcionality hry. Najväčší dôraz sa v nej kladie na korpusový manažér Sketch Engine a jeho jednotlivé nástroje.

Sociálnej sieti Facebook a v nej dostupnému vývojovému prostrediu Facebook Developers sa venuje štvrtá kapitola.

Piata kapitola predstavuje podrobnejšie samotnú hru X-plain a jej funkcionalitu v pôvodnej verzii. V skratke sú v nej vymenované aj existujúce alternatívne projekty.

Šiesta kapitola je pre prácu kľúčová, pretože popisuje návrh hry a najdôležitejšie body implementácie. Venuje sa fázam návrhu s dôrazom

na dôvody výberu jednotlivých technológií a ku koncu kapitoly popisuje riešenia k vzniknutým problémom v priebehu vývoja.

Siedma kapitola sumarizuje splnenie zadania a výsledky práce.

2 GWAP's

V oblasti návrhu počítačových hier dochádza k neutíchajúcemu pokroku. Dôvodom nie je len rapídny rozvoj na poli elektroniky a počítačovej grafiky, ale predovšetkým originálnosť a novátorstvo v námetoch na hry. Dlhodobo sa na hry pozeralo ako na jednostranne zameranú činnosť, ktorej hlavnou úlohou bolo zabaviť a zaujať, no dnes sa stretávame s konceptmi hier, ktoré už majú aj iný primárny účel. Môže to byť napríklad vysvetlenie a priblíženie nejakého javu ako v hre od pracovníkov MIT herného laboratória „Slower speed of Light" [5] alebo využitie potenciálu hry na rozvoj v určitej vedeckej sfére. Práve druhá spomenutá skupina hier je označovaná ako Games with a Purpose.

Za jej autora sa považuje profesor Luis von Ahn, pôsobiaci na univerzite Carnegie Mellon University v Pittsburghu, ktorý nielen teoreticky opísal princíp hier, ale s niekoľkými spolupracovníkmi vytvoril hru ESP Game, ktorá už niekoľko rokov úspešne dokazuje užitočnosť tohto konceptu. Sám autor tento koncept vystihol takto: „V GWAP's ide o to, že prostredníctvom on-line hier môžu ľudia spoločne riešiť rozsiahle výpočtové problémy" [1]. Užívatelia poskytnú materiál hraním počítačovej hry a následne sa stávajú priamymi riešiteľmi určitej úlohy takmer bez vynaloženého úsilia. Samozrejme, aby dáta boli použiteľné, je potreba ich overiť. Dôležitú úlohu pri návrhu zohráva maximalizácia počtu hráčov, pretože princíp hry spočíva v nazbieranom množstve dát.

2.1 Common sense facts

Za cieľovú skupinu dát v koncepte GWAP's sa považujú informácie, ktoré sú získavané z hier a majú aj sekundárne využitie mimo nich. Naša hra sa zameriava na skupinu jazykových dát, konkrétne CSF. Pod týmito faktami si človek môže predstaviť ľubovoľné, jemu samozrejmé, informácie, neformálne označované aj ako „sedliacky rozum". Hovorím o javoch, ktoré z ontologického hľadiska Barry Smith popisuje ako autonómne a existujúce nezávisle na ľuďoch [6]. Ich veľkosť sa postupne s evolúciou zväčšuje, a to,

čo bolo nedávno nové, je dnes považované za CSF. Kremene trením o seba iskria a vzniká oheň.

Najčastejšie sa takéto dáta získavajú pomocou slovných hier zameraných na opis, zoskupovanie alebo triedenie skupín slov. Získavanie CSF, spolu s rozpoznávaním objektov alebo strojovým prekladom sú pre počítače veľmi náročným procesom a priniesť skutočne kvalitné výsledky takýchto úloh je možné len v spolupráci s technológiami, ktoré simulujú ľudské postupy. Danými problémami sa zaoberá oblasť umelej inteligencie, ktorá s pomocou nástrojov spracovania prirodzeného jazyka získava dôveryhodné výstupy. Takto vyextrahované CSF môžu svojou podstatou prispievať k ontológii jazyka alebo byť využiteľné na tvorbu ďalších nástrojov v oblasti umelej inteligencie. Napríklad pri tvorbe vyhľadávacieho nástroja, kde keď sú k dispozícií CSF: „lieky odstránia bolesť", „lieky sú v lekárni" a zadáme na internete „mám bolesti", automaticky sa zobrazia odkazy na dostupné lekárne

Keďže v GWAP's dochádza k náhodnému výberu dvoch spoluhráčov, u ktorých sa neočakáva vzdelanie na rovnakej úrovni, základom úspechu pri hraní sú spoločné asociácie. Práve to zaručuje, že úspech a vysokú početnosť budú mať len tie dáta, ktoré sú dostatočne všeobecné, teda zaraďované k CSF.

K nim sa radia aj informácie o obrázkoch, videách a nahrávkach umiestnených na internete, ktoré sú súčasťou rovnako účelovo zameraných hier. Sú to jedno či viac slovné popisy týchto multimédií od hráčov, ktorí ich takto pomáhajú presnejšie kategorizovať. Zvyšujúca sa kvalita obsahu stránok internetu v dôsledku množstva takýchto nazbieraných popisných dát, vedie k efektívnejšiemu vyhľadávaniu na Internete.

2.2 Vlastnosti dát

Návrh každej z GWAP's je nasmerovaný k zberu kvalitných dát. Ako objasňujú autori práce [7], mali by byť čisté, rozsiahle a užitočné.

Za čisté sa považujú údaje, ktoré sú overené a mám istotu, že sa do hry nedostali podvodom. To sa dá obmedziť tak, že vygenerované dáta označím za dôveryhodné, až keď na ne narazím po niekoľkýkrát a sú nado-

budnuté z viacerých zdrojov. Počas hry nie je umožnená iná komunikácia, ako spoločný obrázok alebo slovo, ktoré obaja hráči vidia a potrebujú sa zhodnúť aspoň v jednom zo slov, ktorými ho opisujú. Ďalším zabezpečením je rozdielna IP adresa, aby hráči nemohli obchádzať pravidlá vzájomným dorozumievaním sa a falšovať výsledky.

Ako som spomenula na začiatku, účelom hry je zhromaždenie čo najrozsiahlejšieho množstva dát. Zaujať čo najviac hráčov je samozrejme cieľom každej hry, no pri GWAP's sú to odlišné a hlavne hlbšie dôvody. Namiesto nadobudnutia finančného zisku či stáleho miesta na trhu, sa snaží o vysoký počet hráčov a aktívnu hernú stránku hlavne kvôli zisku hodnotných dát, ktoré majú potenciál byť využité na vedecké účely. Rozsiahlosť dát zabezpečí zároveň aj ich automatickú korekciu.

Vygenerované dáta musia byť užitočné, tzn. mať uplatnenie pri riešení konkrétneho problému. Táto požiadavka môže byť v konflikte s predchádzajúcimi dvoma, pretože vytvoriť hru, ktorá takéto užitočné dáta generuje, zároveň je atraktívna a ešte obsahuje kvalitné a čisté dáta, je ťažké.

2.3 Gamifikácia[1]

Úloha zaujať širokú skupinu užívateľov a rozširovať jej pôsobnosť aj ďalej je cieľom každého projektu. Prostriedkom na uskutočnenie týchto snáh je gamifikácia, ktorej definíciu uviedli autori práce [8] ako: „používanie herných návrhov v nehernom kontexte". Atraktivita počítačových hier sa tak namiesto zdroja závislosti na nich stáva spôsobom, ako na herných princípoch vybudovať užívateľom prívetivé a úspešné aplikácie.

Ako autori článku vysvetľujú, že nehovoríme o hravosti (*playfulness*), ktorú v hrách reprezentujú prvky ako voľnosť, expresívnosť a búrlivosť, ale zameriavame sa skôr na súťaživý charakter hry (*gaming*), vymedzenie pravidiel, prípadne motivačné odmeny, ktoré sa snažím do aplikácie prebrať.

Prelomovou je v tejto oblasti aplikácia „Foursquare", ktorá zaznamenáva navštívené miesta jej majiteľa, a za splnenie určitých úloh udeľuje výhry a vylepšenia. Rovnako umožňuje užívateľovi podeliť sa o svoje úspechy

[1] pochádza z anglického gamification.

s priateľmi na sociálnej sieti. Aplikácia sa stala fenoménom hlavne preto, že na rozdiel od video hier, ktoré vyžadujú veľké množstvo času stráveného používaním elektroniky, je prostriedkom na dobrodružstvo a akciu v reálnom živote. Po uvedení tejto hry sa gamifikácia stala cieľom mnohých vedeckých štúdií a začala prenikať nielen do marketingovej oblasti, ale aj do rozličných akademických projektov.

Kvôli tomu, že účinnosť GWAP's závisí na ich obľúbenosti a schopnosti zaujať užívateľa, je nevyhnutné, aby spĺňali niektoré z elementov gamifikácie a vytvárali si tak vzťah s užívateľom. Konkrétne odporúčané herné elementy pre danú aplikáciu sa líšia v závislosti na miere vážnosti či cieľoch aplikácie. Herné elementy sú:

- Avatar – užívateľova reprezentácia v hre
- Trojrozmerné prostredie – vytvorenie ilúzie skutočného sveta
- Naratívny kontext – vsadenie do príbehu, kontextu
- Spätná väzba – interakcia s užívateľom
- Reputácia a úrovne – možnosť vylepšiť avatara a postupovať do ďalších úrovní
- Nákupné centrá – využitie nahratých bodov na rozličné vylepšenia ako schopnosti avatara, pomôcky pri hre alebo zmeny rozhrania hry
- Súťaže – príležitosť porovnať svoje sily s viacerými užívateľmi, motivačné prvky pre hráča
- Tímy – vytváranie spoločných stratégií
- Nastavenie paralelných hier – hranie súčasne s priateľmi
- Časové ohraničenie – zvýšenie dynamiky hry

V slovných hrách, akou je X-plain, sa na grafickú stránku nekladie primárny dôraz, preto sú avatar a trojrozmerné prostredie v našom prípade irelevantné. Naopak dôležitosť je pripisovaná časovému ohraničeniu a súťaživosti, ktorá je základnou zložkou hry. Spätná väzba vo forme určitej interakcie s užívateľom je uplatniteľná v každom druhu hry. Ostatné zložky sú možnosťami, ktoré môžu k atraktivite hry len prispieť.

3 Nástroje NLPC

Laboratórium spracovania prirodzeného jazyka (NLPC) Masarykovej univerzity sa zameriava na výskum v oblasti syntézy a rozpoznávania jazyka s využitím najnovších poznatkov v oblasti jazyka aj informatiky. Hlavným cieľom je vďaka kooperácii jazykovedných a počítačových expertov dospieť k výsledkom v oblasti lexikálnych databáz, reprezentácie jazykových výrazov a desambiguácie[2] korpusových dát. Pre označenie slovných tvarov, značiek a konkrétnych príkladov slov budem v nasledujúcich kapitolách používať odlišný druh písma.

3.1 Morfologický analyzátor *majka*

Morfologický analyzátor je nástroj, ktorý dokáže slová v texte lematizovať[3] a vygenerovať pre ne značku, obsahujúcu informácie o jeho gramatických kategóriách. V súvislosti s dátami, ktoré značka reprezentuje, sa rozlišuje v Českej republike brnenský a pražský systém.

Pražský systém, označovaný vďaka spôsobu ukladania informácií ako pozičný, si fixuje jednotlivé pozície v značke s gramatickými kategóriami osobitými pre každý slovný druh. Značka má teda tvar súvislého reťazca, obvykle s 15 pozíciami (napr. `ANNSA-----S----)`.

V brnenskom, nazývanom aj atribútový, je informácia do značky zakódovaná kombináciou atribútu a hodnoty, čo zvyšuje prehľadnosť a prispieva k úspornosti dátového priestoru. Taktiež v nej nezáleží na poradí pozícií a je jednoducho rozšíriteľná (napr. `k1gMnP13`).

NLPC má k dispozícii analyzátor majka, vytvorený RNDr. Pavlom Šmerkom, Ph.D., ktorý ho vo svojej dizertačnej práci [9] popisuje a podrobnejšie vysvetľuje dôvody jeho vzniku. Majka zachováva princípy brnenského systému a formálne nadväzuje na morfologický analyzátor ajka, autorom ktorého je Mgr. Radek Sedláček, Ph.D. Oproti ajke má tento analyzátor niekoľko vylepšení. Jedným z nich je úspornejšia reprezentácia

[2] zjednoznačňovanie korpusových značiek.
[3] previesť slovo do základného tvaru.

s cieľom zabrániť redundancii výrazov v slovníku. Rovnako ako v ajke, aj tu je každé slovo reprezentované ako trojica *word: lemma: tag (WLT)*, no pri zápise sú namiesto celého slova použité veľké písmená označujúce počet písmen, ktoré musíme od slova odtrhnúť (A = 0, B = 1, ...) a reťazec, ktorý k takto orezanému slovu treba pridať.

```
ježek:A:k1gMnSc1
ježka:Cek:k1gMnSc2
```

V tomto zápise je v prvom riadku slovo *ježek* v základnom tvare (A = 0 znamená bez zmeny). V druhom riadku je toto slovo v genitíve a základný tvar vznikne oddelením 2 písmen (C = 2) a pridaním prípony *ek*.

Ďalšou zmenou je úprava značkovania v snahe o štandardizáciu, zjednodušenie a desambiguáciu pri vyhľadávaní [10]. Jednou z nich je napríklad zredukovanie počtu hodnôt atribútu „*D*“ (číslo) z pôvodných štyroch len na dve (jednotné a množné), keďže zvyšné hodnoty (duálne, rodinné) sa dajú chápať ako konkrétnejší druh ponechaných. Naopak pridaný bol atribút pre označenie interpunkcie „*kI*“, a taktiež jeho subklasifikácia „*x*“. Kompletná tabuľka označení analyzátoru majka je uvedená na obrázku 3.1.

Majka na rozdiel od ajky poskytuje osobité nástroje na každú z potrebných operácií. V prípade potreby analyzovať teda analyzátor, získať všetky tvary slova generátor atď. Výhodou je možnosť úpravy jedného z nich bez zmeny ostatných.

Najväčším rozdielom je však návrh analyzátoru, pretože dáta určené na analýzu sú v tomto návrhu konečným zoznamom trojíc WLT a teda na vyhľadávanie nám stačí konečný automat, čo niekoľkonásobne urýchli vyhľadávanie v ňom.

Common attributes

k	Part-of-speech
1	Substantives
2	Adjectives
3	Pronomina
4	Numerals
5	Verbs
6	Adverbs
7	Prepositiona
8	Conjunctions
9	Particles
0	Interjections
I	Punctuation
g	Gender (k1–k4)
M	Animate masculine
I	Inanimate masculine
F	Feminine
N	Neuter
c	Case (k1–k4, k7)
1–7	First–Seventh
n	Number (k1–k4)
S	Singular
P	Plural
e	Negation (k2, k5, k6)
A	Affirmation
N	Negation
d	Degree (k2, k6)
1	Positive
2	Comparative
3	Superlative
p	Person (k3, k5)
1–3	First–Third
w	Stylistic subclassification (k0–k9)
B	Poeticism
H	Conversational
N	Dialectal
R	Rare
Z	Obsolete
z	Common subclassification (k1–k9)
S	Contains -s enclitic
Y	Word form of *aby, kdyby, by*
A	Abbreviation

⋆ ~	Statistical characteristics
0–9	E. g. frequency

k1 subclassification

⋆ x	Type
P	Word form of *půl*
F	Family surname

k3 subclassification

x	Type
P	Personal
O	Possessive
D	Demonstrative
T	Delimitative
y	Type
F	Reflexive
Q	Interrogative
R	Relative
N	Negative
I	Indeterminate

k4 subclassification

x	Type
C	Cardinal
O	Ordinal
R	Reproductive
y	Type
N	Negative
I	Indeterminate

k5 subclassification

m	Type (mode)
F	Infinitive
I	Present Indicative
R	Imperative
A	Active part. (past)
N	Passive part.
S	Adv. part. (present)
D	Adv. part. (past)
B	Futreu indicative
a	Aspect
P	Perfect
I	Imperfect

k6 subclassification

x	Type
D	Demonstrative
T	Delimitative
y	Type
Q	Interrogative
R	Relative
N	Negation
I	Indeterminate

⋆ t	type
S	Status
D	Modal
T	Expresses time
A	Expresses respect
C	Expresses reason
L	Expresses place
M	Expresses manner
Q	Expresses extent

k8 subclassification

x	Type
C	Coordinate
S	Subordinate

kI subclassification

x	punctuation list
.	.!?
,	,:;
"	"'‘,,”
(	[{L<
)	]}]>
~	~$%^&-_+=|/# etc.

Attribute-to-PoS assignment

k	attributes
1	gnczw~
2	egncdzw~
3	gncpxyzw~
4	gncxyzw~
5	eampgnzw~
6	edxytzw~
7	cw~
8	xzw~
9	zw~
0	w~
1	x~

Canonical ordering

kegncpamdxytzw~

Obr. 3.1 Tabuľka značiek analyzátoru majka [10].

3.2 Značkované korpusy

Korpus je elektronická alebo písaná databáza jazykových prvkov vo forme súvislého textu.

Delenie:

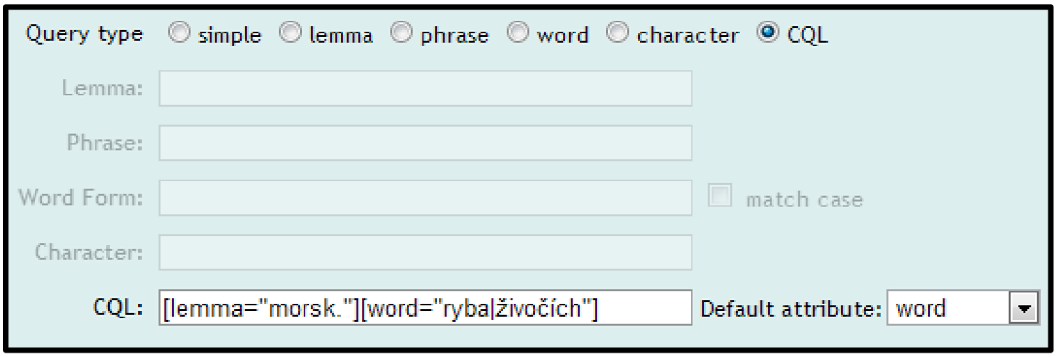
- podľa zdroja:
    - korpus písaného textu – písaný zdroj

    - korpus hovoreného textu – zvukový zdroj
- podľa jazyka:
    - jednojazyčné – v jednom jazyku
    - paralelné – súbežne v dvoch jazykoch s ohľadom na zachovanie súbežnej formy oboch textov
- podľa zamerania:
    - všeobecné – bez špeciálnych požiadaviek
    - reprezentatívne – obsahujúce dostatočne široké spektrum slov zo zadaného špecifického žánru
- podľa uloženia:
    - v textovej forme – jednoduchý zdrojový text
    - značkovanej forme – každý z prvkov textu má uloženú množinu vlastností

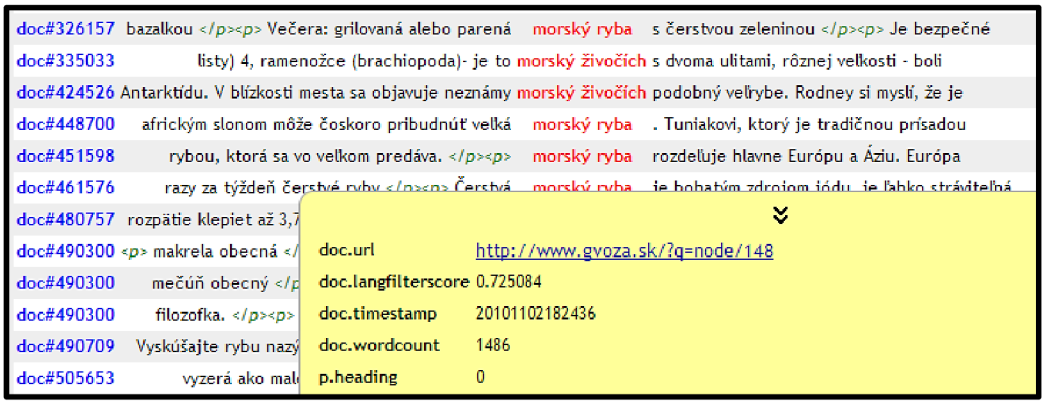
Naša hra pracuje s písanými, českými korpusmi všeobecného typu, ktoré sú značkované. Proces pretvorenia textového korpusu na jeho značkovanú formu použiteľnú pre účely spracovania prirodzeného jazyka zabezpečuje práve morfologický analyzátor majka (pozri skoršie v kapitole 3.1)

Na prácu s korpusmi sú k dispozícii odkazovacie jazyky špecializované na vyhľadávanie v rozsiahlych textových súboroch a umožňujúce aj profesionálnejší prístup. V prípade NLPC nám na prácu s takto označkovaným korpusom slúži korpusový dotazovací jazyk (*Corpus Query Language CQL*), pomocou ktorého môžeme vyhľadávať nielen slová, ale aj celé vetné konštrukcie zadané pomocou regulárnych výrazov tohto jazyka. Príkladom je použitie napríklad znaku „.“ pre výber výrazu obsahujúceho na mieste bodky akékoľvek písmeno alebo znak, alebo „|“ pre vyjadrenie „alebo“. Názorná ukážka použitia je na obrázku 3.2.

Výhodou CQL je možnosť vytvárať výrazy postupne, teda na začiatku si vyformovať základné výrazy, a tie potom postupne „obaľovať“ ďalšími až na finálnu podobu výrazu. Takto spracovaný korpus slúži ako materiál pri tvorbe slovníkov, korektorov, alebo pre naše účely kategorizácie skupín slov.



Obr. 3.2 a Príklad použitia vlastného vyhľadávania v korpusoch pomocou nástroja Concordance a jazyka CQL. Cieľovými sú výrazy obsahujúce slovo „morsk“ v texte nasledované buď slovom „ryba“ alebo „živočích“.



Obr. 3.2 b Výsledok dotazu: „[lemma="morsk."][word="ryba|živočích"]“ z predchádzajúceho obrázku. Po kliknutí na modrú položku sa zobrazí originálny zdroj daného korpusu s bližšími informáciami o ňom(žltý rámec.)

3.3 Sketch Engine

Sketch Engine [11] je korpusový manažér NLPC, ktorý prostredníctvom jazyka CQL umožňuje prácu s rozsiahlymi korpusovými dátami. Svojím užívateľom ich poskytuje v značkovanom formáte, a s možnosťou výberu zo širokej škály jazykov. Ponúka rôznorodé jazykové nástroje, z ktorých bližšie popíšem dva najdôležitejšie, ktoré som vo svojej práci aj využila.

3.3.1 Concordance

Nástroj Concordance umožňuje vyhľadať v danom korpuse buď konkrétne slovo, alebo použitím príkazov a konštrukcií jazyka CQL aj celé výrazy so spresnenými požiadavkami (viď skoršie Obr.3.2 a). Takto môžem v texte

vyhľadávať napríklad slovné spojenia so špecifikovanými predponami či príponami alebo slovné spojenia nachádzajúce sa v texte bezprostredne za sebou. Po vygenerovaní zoznamu konkordancií[4] tento nástroj ponúka aj ďalšie funkcie, ako napr. pozitívny a negatívny filter.

3.3.2 Word Sketch

Word Sketch je hlavným nástrojom aplikácie Sketch Engine a jeho účelom je pre zadané slovo nájsť v korpuse najfrekventovanejšie slová spojené so zadaným určitou väzbou. Tieto slová sa zobrazujú v stĺpcoch, roztriedené podľa druhu relácie, ktorou sú spojené so zadaným slovom. Relácie delíme do 3 skupín:

- symetrické - *coord*
  - spájajú slová v koordinácii, vzájomne zameniteľné
  - slová označujú rovnaký slovný druh
  - spojene spojkami a, alebo
  -napr. pre slovo „pes“ je v stĺpci *coord* „mačka“
- duálne - *modifies, a_modifier*
  - *gen_1, gen_2*
  - vždy dve relácie, ktoré sú vzájomne inverzné
  - napr. pre slovo „mozog“ je *a_modifier* „ľudský“ a pre „ľudský“ je *modifies „mozog“*
- ternárne - *prec_<predložka>*
  - *post<predložka>*
  - dvojica slov, spojená predložkou uvedenou v názve relácie
  - *post* znamená, že zadané slovo je za predložkou, *prec* že je pred ňou
  - napr. pre slovo „auto“ je *post_na* „parkovisko“, predlohou z textu pre túto reláciu je spojenie „auto na parkovisku“

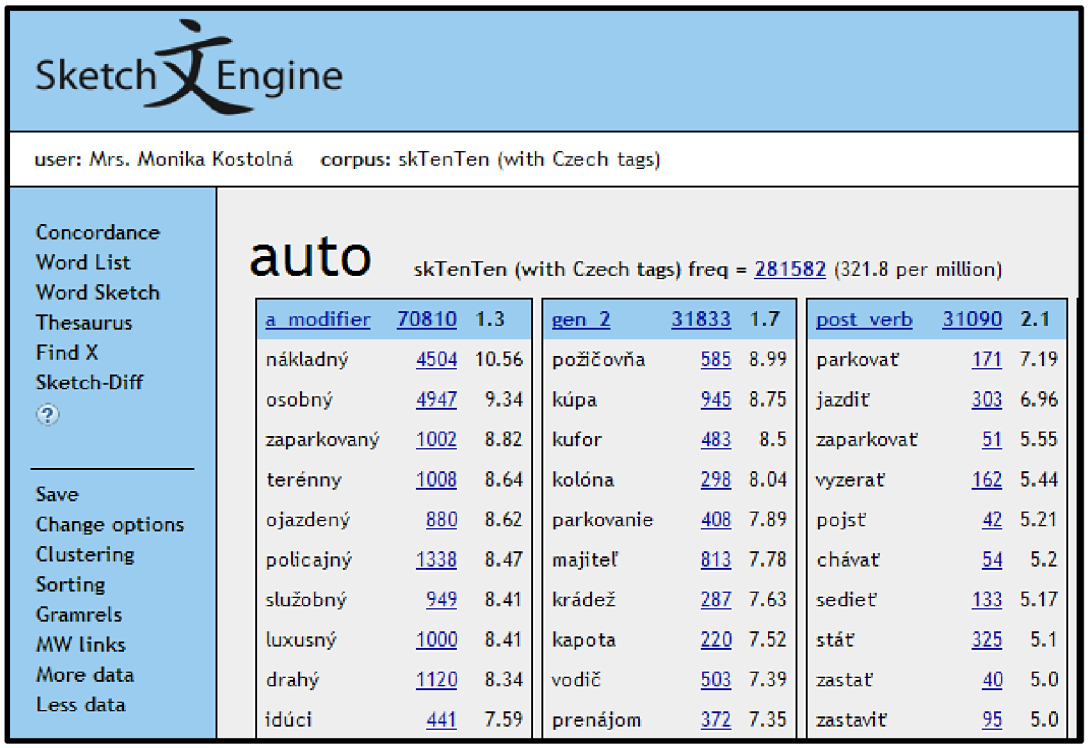
Každé zo slov má okrem príslušnosti k relácii uvedený aj počet výskytov v celom korpuse, skóre udávajúce ustálenosť daného spojenia vzhľadom

[4] interaktívny odkaz na originálny text.

k ostatným výskytom tohto slova a odkaz na kontext v korpuse. Obrázok 3.3 zobrazuje výsledok pre slovo „auto“.

Pri vyhľadávaní mám možnosť využiť aj rozšírené možnosti s dôrazom na kvantitatívne alebo kvalitatívne vlastnosti dát. Ako jeden z parametrov pri rozšírenom vyhľadávaní je napríklad zobrazenie iba tých slov, ktoré majú zadaný minimálny počet výskytov alebo výšku skóre. Prístupný je aj náhľad do použitého kódovania v jazyku CQL, užitočný pre pokročilejších užívateľov.

Každý z nástrojov pracuje po celú dobu s jedným, nami zvoleným, jednojazyčným korpusom. V tejto práci som používala dva pôvodné: „czes“ a „sktenten“, a modifikáciu druhého spomenutého „sktenten_cstag“ (viac v kapitole 6).



Obr. 3.3 Ukážka Word Sketch relácií pre slovo auto. Podčiarknutá položka udáva počet výskytov v korpuse, po kliknutí na ňu sa zobrazí v kontexte. Čierna, číselná položka udáva skóre (ustálenosť). [12]

4 Dostupné programovacie prostredie Facebook API

Webová sociálna sieť Facebook je primárne určená na sociálnu interakciu medzi ľuďmi a v súčasnej dobe má už vyše 1,1 miliardy používateľov [13]. Od roku 2007 je v nej voľne dostupné programátorské prostredie Facebook Developers, ktoré ponúka svojím užívateľom aj možnosť vytvárania rozličných aplikácií. Alternatívne sociálne portály ako MySpace alebo LinkedIn, predstavujúce jej najväčších konkurentov, žiadnu obdobnú funkcionalitu neobsahujú.

4.1 Facebook Developers

Facebook Developers je súčasťou sociálnej siete Facebook a umožňuje externým webovým aplikáciám fungovať v jeho prostredí bez potreby presmerovania a prídavných modulov. Programátorom ponúka možnosť vytvorenia sieťových aplikácií alebo aplikácií určených pre mobilné zariadenia aj s vstavanou integráciou a synchronizáciou. Po spustení aplikácií ostávajú užívatelia naďalej prihlásení, zdieľajú grafické rozhranie siete a zároveň majú k dispozícii kompletnú funkcionalitu. Tento nástroj je vhodným prostriedkom na propagáciu hry medzi čo najširšiu verejnosť.

5 Hra X-plain – pôvodná verzia

Prvá verzia hry X-plain bola vytvorená členmi NLPC Masarykovej univerzity. Podľa svojho charakteru je zaradená medzi slovné a v pôvodnej verzii je to výlučne česká hra pre jedného užívateľa, kde úlohu spoluhráča zastupuje automatizovaný hráč (ďalej bot). Bot je vytvorený tak, aby čo najvernejšie simuloval ťahy človeka.

5.1 Bot

Herný koncept počítača ako bota, ktorý je na nerozoznanie od hráča- človeka nie je ničím novým, vo svojej práci [7], ho trojica profesorov z Carnegie Mellon University popisuje ako náhradu reálneho hráča v alternatíve ku klasickej hre Verbosity. Potreba vytvoriť takúto verziu sa u nich objavila pri problémoch s nečakaným odpojením spoluhráča, kedy bolo prerušené spojenie. Vďaka množstvu dát mohli vytvoriť bota, určeného na okamžité zastúpenie role druhého hráča.

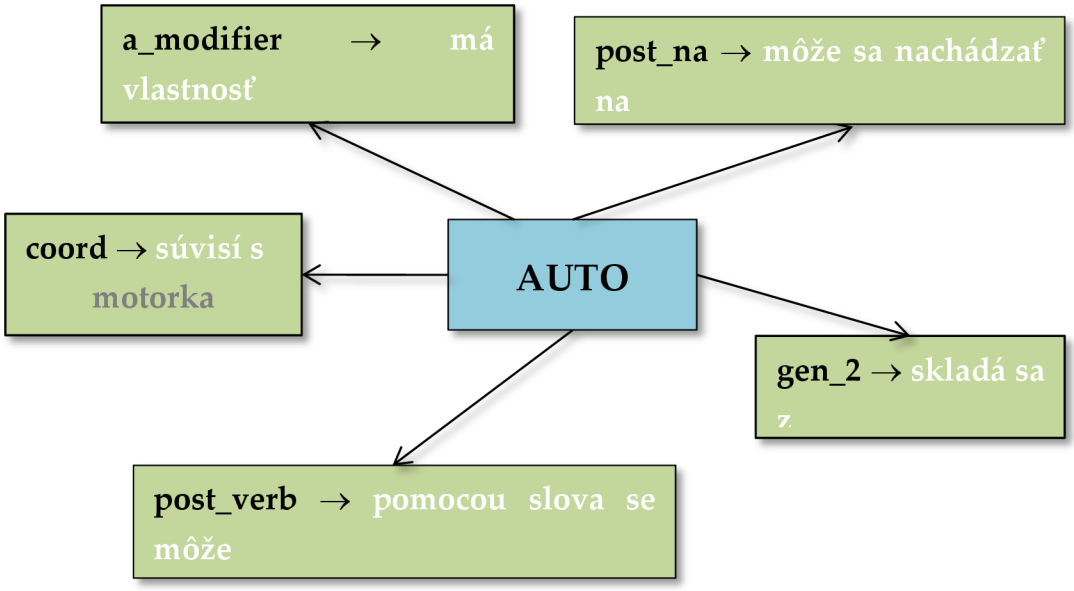
5.2 Použité nástroje

Prostriedkami, ktoré našej hre umožňujú používať bota, sú nástroj Word Sketch a databáza našej aplikácie obsahujúca štruktúrovanú históriu hier. V našom prípade je nasadenie bota nevyhnutné, pretože hra zatiaľ nemá toľko používateľov a šanca, že daný hráč bude mať v ľubovoľný čas k dispozícii spoluhráča, je veľmi malá.

Česká verzia používa pre úpravu vstupných dát od užívateľa morfologický analyzátor majka, ktorý zadaný výraz upraví na základný tvar a v takejto podobe môže byť vyhľadaný vo Word Sketch. Rovnako pri vyberaní tipov, ktoré v hre použije bot, sú informácie vyextrahované z Word Sketch a prevedené zo základného tvaru do odpovedajúceho pádu príslušnej relácie.

5.3 Cieľ hry

Cieľom hráča je kooperatívnym hraním získať čo najviac bodov. Na začiatku sa hráčovi zobrazí tajné slovo, ktoré by mal pomocou niekoľkých ponúknutých relácií opísať spoluhráčovi (botovi). Ak ten presne toto slovo uhádne, úlohy sa vymenia a daný hráč následne čaká na indície od bota, na základe ktorých má hádať. Pokiaľ v ľubovoľnej časti hry nevie, má možnosť vzdať sa a pokračovať opäť výmenou hráčov a vygenerovaním nového slova. V hre je trojminútový časový limit, ktorý vynucuje rýchlu odozvu. Hráč má možnosť si v rekapitulácii hry prehliadnuť, ako svoju poslednú, tak aj ľubovoľnú staršiu hru, spolu s použitými tipmi a indíciami. V štatistikách tvrdení sa mu k ľubovoľnému zadanému slovu zobrazia všetky trojice:



Obr. 5.2 V modrom obdĺžniku zadané slovo. V zelených obdĺžnikoch je ukážka prekladu relácií zo sketcha (čierna farba písma) do sémantického významu použiteľného v našej hre (biela farba) s konkrétnym príkladom slova z tejto relácie (sivá farba).

`subjekt - predikát - objekt`, v ktorých sa toto slovo nachádza.

5.4 Preklad relácií *„z Gramatiky do sémantiky“*

Databáza využívaná vo Word Sketch je vytvorená pomocou morfologickej analýzy, ktorá je založená predovšetkým na gramatických štruktúrach. V našej hre ide o zber slovných spojení, ktoré majú významový súvis, preto

je k použitým reláciám z Word Sketch vytvorený ekvivalentný preklad *„z gramatiky do sémantiky"*[4]. Využívajú sa len tie relácie, u ktorých je takýto preklad zmysluplný, a z nich je v každom kole náhodne vybraná časť, ktorá sa bude zobrazovať užívateľovi. Príklad prevodu niekoľkých relácií je naznačený na obrázku 5.2.

Podobný postup sa uplatňuje aj opačným smerom pre prevod *zo sémantiky do gramatiky.*

5.5 Súvisiace projekty

5.5.1 Vedecké práce

- **CyC[14]** – Jeden z prvých projektov v oblasti zberu CSF. Platení experti v ňom zhromaždili vyše milióna výrokov vo forme logického výrazu, za pomoci logického jazyku CycL odvodeného od Lispu.
- **OpenMind** –Webová aplikácia zbierajúca relácie medzi slovami prostredníctvom dobrovoľníkov. Informácie z nej sú časťou ConceptNet5 [15], medzinárodnej sémantickej siete CSF.

5.5.2 Vedecko-herné koncepty

- **ESP Game a Verbosity[3]** –Koncepty od zakladateľa GWAP's Luisa von Ahna, ktoré učinili najväčší prelom v oblasti hier zbierajúcich CSF. Hra ESP Game je hra pre dvoch hráčov, ktorej cieľom je opísať súbežne videné vizuálne dáta a dosiahnuť so spoluhráčom zhodu aspoň v jednom zapísanom slove. V hre Verbosity si hráči podobne ako v X-plain striedajú úlohy hádania a napovedania slova. Oproti X-plain sú obe hry založené na interaktivite medzi dvoma reálnymi hráčmi. Verbosity má podobne ako X-plain k dispozícii bota nahrádzajúceho úlohu spoluhráča, ale tento sa zapája do hry len v 5 percentách všetkých hier.
- **Akinátor [16]** –Viac-jazyčná hra, v ktorej sa bot, v hre znázornený ako duch z lampy, pokúša zistiť meno významnej osoby tajne vybranej hráčom. K menu osoby sa snaží dopracovať pomocou kladenia otázok o subjekte s možnosťou výberu odpovede rôznych úrovní áno

a nie. Hra sa rovnako ako X-plain zdokonaľuje každým ťahom a hlavnú úlohu v nej zohráva rozsiahla databáza, v tomto prípade s veľmi poprepájanými dátami, ktorá je vďaka úrovňovej štruktúre odpovedí takmer nepochybyteľná.

- **Guess My Word [17]** –Anglická hra na sociálnej sieti Facebook, v ktorej sa dvaja hráči medzi sebou striedajú v napovedaní a hádaní. Na rozdiel od hry X-plain tu majú možnosť výberu z 3 ponúkaných slov a pri hádaní je obmedzený počet ponúkaných písmen.

5.6 Splnenie vlastností GWAP's

Ako som spomínala skoršie (podkapitola 2.2), hra by mala splňovať trojicu vlastností identifikujúcich GWAP's hry.

Keďže automatizovaný hráč, ktorého hra používa, čerpá svoje ťahy z databázy plnej dát reálnych hráčov, jeho indície a tipy sú považované za tipy človeka. Šanca uhádnuť ťah je teda závislá od miery podobného zmýšľania aktuálneho hráča a autora údaju z databázy, a priamoúmerne stúpa s všeobecnosťou a bežnosťou zadaných dát. Takéto dáta už je možné označiť ako užitočné.

Rozsah českých dát je v pôvodnej verzii hry nemalý, no vylepšenia, ktoré sú cieľom tejto práce, by mali priniesť ich nárast, väčšiu kvalitu a rozšírenie o slovenskú množinu.

V otázke čistoty som sa stretla s niekoľkými problémami, ktoré sú zapríčinené predovšetkým nedbanlivosťou užívateľov alebo neprítomnosťou cudzojazyčných verzií. Z nich napríklad uvádzanie výrazov bez diakritiky, slov v nesprávnom páde, čísle alebo v inom jazyku (zvyčajne v slovenčine). Väčšiu časť týchto problémov je však možné zanedbať, pretože aj pri používaní nástrojov, s ktorými pracujem, dochádzalo k porušovaniu noriem ich užívateľmi , a za niekoľko rokov sa aj k takýmto slovám nazbieralo dostatočné množstvo dát. Problém so slovami v slovenskom jazyku vyriešim pridaním verzie pre slovenčinu.

6 Rozšírenie hry X-plain

6.1 Terminológia a značenie

Názvy premenných, funkcií, tried a iných prvkov implementácie budem značiť kurzívou. Odlišný druh písma použijem rovnako ako v kapitole 3 pre konkrétne príklady slov alebo značiek. Ďalej nižšie uvádzam niekoľko termínov, ktoré budem kvôli ich vysokej frekvencii v tejto časti práce uvádzať v skrátenom tvare.

- Nástroj Sketch Engine → SkE
- Počítačový hráč (ComputerPlayer) → bot
- Ľudský hráč (HumanPlayer) → človek

6.2 Objektový návrh hry

Samotná hra je navrhnutá objektovo pomocou jazyka PHP, a jej jednotlivé triedy sú: *Game*, *Round*, *ComputerPlayer*, *HumanPlayer*, obe posledné implementujúce abstraktnú triedu *Player*.[5] Program počas hry vytvára jednotlivé objekty týchto tried. Skript v PHP je počas hry opakovane spúšťaný pre rôzne požiadavky od užívateľa. Aby program o objektovú reprezentáciu hry neprichádzal, využívam v hre na ukladanie kompletnej informácie o prebiehajúcej hre a aktuálnych objektoch *session* (v preklade *sedenie*). Objekty sú v ňom pomocou serializácie uložené ako reťazce a takto sú uchované medzi jednotlivými spusteniami skriptu na serveri.

6.2.1 Trieda *Game*

Trieda *Game* predstavuje hru s odkazmi na tieto objekty:

- *HumanPlayer*
- *ComputerPlayer*
- Aktuálne *Round*

[5] Názvy tried v preklade: hra, kolo, bot, ľudský hráč, hráč

Do prvého kola vstupujú premenné odkazujúce na hráčov vždy identicky naplnené, teda človek v úlohe napovedajúceho a bot v úlohe hádajúceho. Po skončení každého kola sa obsah premenných vymení (zámena úloh hráčov).

Game si udržiava premennú obsahujúcu čas začiatku hry, ktorá obmedzuje časový limit hry na 3 minúty, a premennú na počítanie súčtu nahratých bodov z celej hry. Pôvodný výpočet skóre bol založený na udeľovaní bodov vzhľadom na uplynutý čas od začiatku hry. Tento spôsob som nahradila bodovaním, kde počet získaných bodov klesá s použitím každej ďalšej indície. Za každé slovo môže človek získať maximálne 1000 bodov. V závislosti na hráčovej aktuálnej úlohe sa táto suma delí buď počtom ním zadaných indícií, alebo počtom indícií, ktoré si vyžiadal od bota. Vzorec na výpočet skóre: *2000 / (1 + počet použitých indícií v kole).*

6.2.2 Trieda *Round*

Táto trieda reprezentuje jedno kolo hry, v rámci ktorého rozlišujeme 4 fázy:

- Začiatok
- Získanie indícií
- Získanie tipov
- Koniec

Začiatok predstavuje vygenerovanie tajného slova. Nasledujúce dve fázy sa striedajú tak dlho, kým človek neuhádne tajné slovo alebo nezadá požiadavku „Neviem, ďalšie slovo“. Vtedy sa kolo presunie do fázy koniec, a program o tom informuje riadiacu triedu *Game*, ktorá pripočíta skóre, vymení hráčov, zvýši počítadlo počtu kôl a začne ďalšie kolo.

Funkčnosť programu vo fázach získania tipov a indícií je závislá na druhu hráča, ktorý je na rade a vstupných dátach. Všetky prípustné kombinácie sú popísané v tabuľkách Tab. 6.2.2a a 6.2.2b. Pri ťahoch bota sa na vstupe nenachádza nič, pretože vstupné dáta prichádzajú iba od človeka metódou post. Bot sa teda rozhoduje jedine na základe aktuálnej fázy.

Tab. 6.2.2a Fáza získania indícií

vstup	na rade	akcia programu	ďalšia fáza
vzdanie sa	človek	-	koniec
-	človek	vytvorí pre človeka formuláre na zadanie indícií	získanie indícií
	bot	získa indície od bota	získanie tipov
indícia	človek	úprava indície do nominatívu uloženie indície do databázy	získanie tipov

Tab. 6.2.2b Fáza získania tipov

vstup	na rade	akcia programu		ďalšia fáza
vzdanie sa	človek	-		koniec
-	človek	vytvorí pre človeka formuláre na zadanie tipov a zobrazí jednu indíciu		získanie tipov
	bot	načítanie tipov od bota a ich vyhodnotenie	jeden tip správny	koniec
			všetky tipy nesprávne	získanie indícií
tip	človek	vyhodnotenie tipu	správny tip	koniec
			nesprávny tip	získanie tipov
požiadavka o ďalšiu indíciu	človek	vráti ďalšiu zo získaných indícií		získanie tipov

6.2.3 Trieda *Player*

Triedy *ComputerPlayer* a *HumanPlayer* spracovávajú rovnaké metódy, ale každá rozdielnym spôsobom, preto využívam abstraktnú triedu *Player,* ktorá reprezentuje všeobecného hráča. Okrem abstraktných metód implementuje metódu pre obe podtriedy, náhodne generujúcu relácie, ktoré sú potom používané v hre. Náhoda pri výbere relácií je ako v novej, tak aj v starej verzii, no rozšírením oproti starej je rozdelenie relácií na dve množiny „dôležité“ a „ostatné“. Toto rozdelenie som zaviedla z dôvodu, že úlohu bota zabezpečuje program prostredníctvom výberu tipov a indícií zo SkE. Ten v reláciách obsahuje aj slová, ktoré nedávajú v spojení so zadaným slovom

zmysluplné výsledky. Na základe tohto sme relácie, ktorých slová korešpondujú s významom zaradili k „dôležité“ a ostatné k „ostatné“. Pri výbere program uprednostňuje množinu „dôležité“, no ich počet dopĺňam aj reláciami zo skupiny „ostatné“.

6.2.4 Triedy *ComputerPlayer* a *HumanPlayer*

Funkčnosť bota program simuluje v triede ComputerPlayer. V častiach hry, kde človek očakáva odozvu od spoluhráča, sa program buď odkazuje do databázy, alebo sa pripája na SkE. Uprednostňuje výskyty v databáze a až v prípade, že v databáze nie je dostatočné množstvo dát, spúšťa script v jazyku Python a pripája sa na SkE. Odtiaľ získava pre zadané slovo množiny slov, spojené s týmto reláciou. Načítané dáta ukladá do výstupného reťazca, ktorý potom používa v hre a zároveň ich ukladá do súborov s príponou *sketch*. Keď sa na slovo opätovne odkazuje a daný súbor nie je starší ako rok, pracuje rovno s ním. Naopak pokiaľ je starší, pripojí sa program na SkE a aktualizuje dáta v ňom. Motívmi na vytvorenie súborov bolo obmedzenie častého pripájania na SkE. Ich vek je overovaný z dôvodu možných úprav a údržby korpusu.

Oproti staršej verzii, kde sa získavali k zadanému slovu len množiny slov z určitých relácií, teraz načítavam zo všetkých, pretože réžia na toto navýšenie je zanedbateľná. Výhodou je, že v prípade pridania ďalších jazykových modulov alebo v prípade doplnení a zmenách relácií je možnosť využiť pôvodné súbory bez potreby načítania nových.

HumanPlayer je trieda vytvárajúca formuláre v HTML pre človeka.

6.3 Slovenská verzia

Program rozširujem o funkčnosť v slovenčine a o možnosť svojvoľného výberu medzi slovenskou a českou verziou. Používanie konkrétneho jazyka je zabezpečené vložením príslušného konfiguračného súboru podľa používaného lokálu do každého súboru, ktorý si prácu s jazykmi vyžaduje (predvolene český cs_CZ).

Pre zabezpečenie funkčnosti slovenskej verzie bol pôvodne v korpusovom manažéri SkE k dispozícii značkovaný slovenský korpus

skTenTen, obsahujúci asi 716 miliónov slovenských pozícií. Oproti značeniu v českom korpuse *czes* v ňom však boli podstatné rozdiely v názvoch relácií a celkovej štruktúre, ktorá nebola dostatočná pre tieto účely. Preto bol použitý jeden z nástrojov NLPC určený na prevod ľubovoľného korpusového značenia do formátu, ktorý je spracovateľný českým analyzátorom. Slovenské relácie z korpusu *sktenten* boli teda zmenené a sprístupnené v novom korpuse: *skTenTen (s českými značkami)*, ktorý v hre aj používam.

6.3.1 Pridanie ďalšieho jazyka

Po dohode s vedúcou práce sa moja úloha rozšírila aj o vytvorenie programu tak, aby sa dal jednoducho rozšíriť o ľubovoľný ďalší jazyk. V prípade potreby rozšírenia teda nie je vyžadované meniť celý kód, ale v niekoľkých bodoch upraviť určité časti programu:

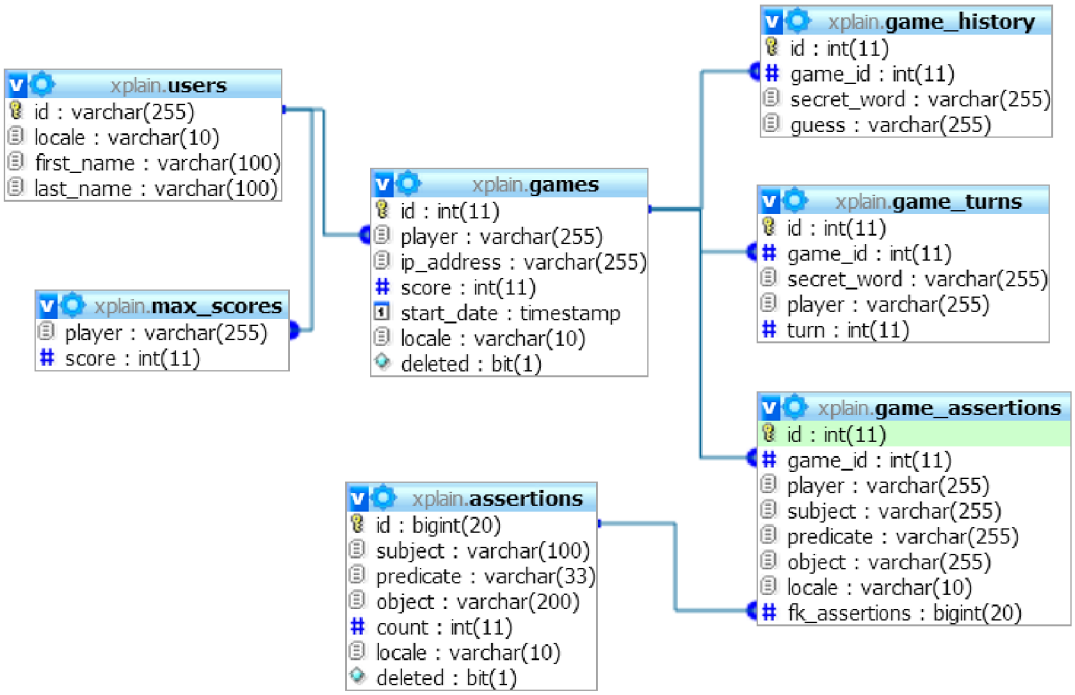
- V súbore *sketch.py* pridať lokál ako jednu z možností a vytvoriť pripojenie na korpus v tomto jazyku.
- Pridať súbor *conf_(nový lokál).inc* a v ňom doplniť preklady grafického rozhrania a relácií z korpusu (presne podľa slovenského alebo českého ekvivalentu)
- Zahrnúť daný lokál medzi podporované jazyky v súbore *conf.inc*.
- Vytvoriť možnosť zmeny jazyka v grafickom rozhraní hry (pridanie vlajky v súbore *menu.inc*)

6.4 Inovácie databázy

Databáza je určená predovšetkým na uchovávanie a správu CSF dát zhromaždených z našej hry, ktoré program spolu s pripojením na SkE využíva na simuláciu úlohy bota. V pôvodnom návrhu nie je kladený dôraz na to, ktorým z týchto dvoch spôsobov dáta získať. V rozšírenej verzii ale program databázu uprednostňuje, a to z dvoch dôvodov. V prvom rade, kvôli väčšej rýchlosti, keďže nie je potrebné pripájať sa na internetovú aplikáciu, za druhé, kvôli predpokladu, že údaje z predchádzajúcich hier užívateľov majú väčšiu šancu uhádnutia. Až keď sa v databáze nenachádza dostatočný

počet záznamov k danému slovu, prechádza program na fázu získania dát zo SkE.

S cieľom vytvoriť možnosť vyhľadať v tabuľke *assertions* trojice subject – predicate – object pre konkrétneho užívateľa, som v tabuľke *game_assertions* vytvorila ku každej z trojíc odkaz na jej kópiu do assertions. Takto je možné identifikovať užívateľa, ktorý by úmyselne vkladal do databázy nepravdivé dáta. Na obrázku 6.4 je znázornený diagram databázy.



Obr. 6.4 Diagram databázy

Vďaka pripojeniu na Facebook mohol program začať do premennej `player` namiesto predchádzajúcej konštanty zapisovať unikátny identifikátor, ďalej v databáze pribudla tabuľka *users* a pohľad *max_score* a vznikli tieto nové atribúty:

- `locale` - jazyk aktuálnej hry
- `first_name` - meno
- `last_name` - priezvisko
- `start_date` - dátum a čas začiatku hry

S modifikáciou atribútu `player`, ktorý jedinečne identifikuje hráča som vytvorila aj pohľad „*max_score*“ s primárnym kľúčom odkazujúcim do tabuľky *users*. Tento pohľad je generovaný rovno v čase potreby z tabuľky *games* výberom vždy najvyššieho skóre z každého celku zoskupeného podľa atribútu užívateľa. Pohľad vytvára dvojice hráč – najvyššie skóre. Použitie generovania pohľadu namiesto statickej tabuľky zabezpečí konzistentnosť, pretože v prípade úprav starších dát alebo zmien v návrhu zdrojovej tabuľky sa automaticky budú meniť aj dáta v pohľade.

6.5 Využitie Facebook API

Cieľom tejto práce bolo vylepšiť hru X-plain a prepojiť ju so sociálnou sieťou Facebook. Prostriedkom na umiestnenie hry na tomto serveri sa stalo rozhranie Facebook API určené na vývoj programovacích aplikácií Facebook Developers nachádzajúce sa priamo v užívateľskom prostredí siete. Toto prostredie je dostupné len užívateľom, ktorí majú vytvorený vlastný osobný profil na danej sieti. Ako užitočný nástroj na získavanie potrebných dát mi slúžil Facebook PHP SDK a Facebook JavaScript SDK. Na rozdiel od bežných skriptov, ktoré si program načítava spolu so stránkou, je JavaScript SDK načítavaný asynchrónne.

6.5.1 Spracovanie osobných údajov

Funkcie poskytnuté spomínanými sú napríklad získanie oprávnení pre používanie osobných údajov užívateľa, zverejnenie výsledkov hry na jeho profile alebo možnosť dať do pozornosti hru aj svojim priateľom. Každá z nich je navrhnutá tak, aby bez povolenia užívateľa neuverejňovala nevyžiadané správy.

Vďaka pripojeniu na Facebook môžem do databázy ukladať aj informácie ako meno, priezvisko a lokál. Osobné údaje slúžia ako na vytvorenie štatistík o priebehu hry, tak aj na filtrovanie tabuliek od nevyžiadaného obsahu alebo možnosť generovať len dáta vybraného užívateľa.

V záujme užívateľov je umožnené zmeniť svoj lokál na ľubovoľný iný z podporovaných a tento príznak o momentálnom jazyku je vždy zaznamenaný v databáze. Pokiaľ sa užívateľ rozhodne zmeniť jazyk, po sú-

hlase s prerušením hry, je automaticky presmerovaný na úvodnú stránku už vo vybranom jazyku.

6.5.2 Hra mimo Facebook

V pôvodnej verzii bola hra prístupná všetkým užívateľom Internetu, preto mohlo umiestnenie na Facebooku spôsobiť stratu tých užívateľov, ktorí nemajú vytvorený profil na tejto sieti. K verzii hry na sociálnej sieti Facebooku som teda vytvorila aj hru mimo nej, aktuálne prístupnú z adresára projektov NLPC. Kvôli nemožnosti jednoznačnej identifikácie užívateľa je ukrátená o niektoré rozšírenia ako prehľad hier užívateľa, porovnanie s priateľmi alebo najlepší hráči. Pokiaľ by užívateľ vyžadoval plnú verziu, má možnosť sa k nej dostať odkazom na Facebook.

6.6 Bezpečnosť aplikácie

V záujme ochrany osobných dát užívateľov je databáza prístupná len administrátorom aplikácie, teda mne a vedúcej mojej práce. Hra zaručuje ochranu osobných údajov užívateľov podľa českého Zákona č. 101/2000 Sb a zaručuje anonymitu v odkaze na zásady súkromia. Cieľom spracovania sú len vstupné dáta a informácie z hry, a rovnako štatistické nástroje sociálnej siete sú založené len na sumách a frekvencii hier. Z druhej strany, hra zabraňuje aj nepovolenému používaniu aplikácie na komerčné účely a odkazuje na licenčnú zmluvu NLPC.

V spojitosti s bezpečnosťou som umiestnila súbor s prihlasovacími údajmi k databáze do chráneného priečinku, odkiaľ má k nemu prístup len správca.

V častiach hry, kde užívateľ zadáva svoje tipy a indície do formulárov, by mohlo v krajných prípadoch dôjsť k *SQL Injections* (v preklade *poškodenia)*. Pre zabránenie užívateľovho zadania takýchto príkazov, používam špeciálne príkazy meniace zadané reťazce znakov na bezpečné pre prácu s databázou.

6.7 Použitie herných elementov Gamifikácie

V kapitole 2.3 je načrtnutých niekoľko herných elementov, ktoré by použitím v hrách či aplikáciách mali viesť k ich väčšej príťažlivosti. Niekoľko z nich sa nachádza aj v našej hre.

- Časové ohraničenie – časový limit 3 minúty stupňuje napätie v hre, a je mierou šikovnosti užívateľa koľko slov stihne uhádnuť.
- Súťaživosť – tabuľky najlepších hier a najlepších hráčov motivujú užívateľov k opätovnému hraniu.
- Spätná väzba – užívateľ môže rozhodovať o počte indícií, ktoré využije k uhádnutiu slova.
- Naratívny kontext – striedanie úloh hráčov, popisovanie tajného slova.

6.8 Implementácia prezentácie dát z hry

6.8.1 Štatistiky a tabuľky výsledkov

Príslušnú databázu k hre je okrem ukladania priebehu hier a základných údajov o užívateľovi možné využiť aj na zatraktívnenie hry. Ponúkajúce sa možnosti pre tvorbu súhrnov a porovnaní sú výsledkom rozšírenia hry o pripojenie k sociálnej sieti, pretože programu poskytujú možnosť triediť dáta podľa rôznych kritérií.

Z pôvodného návrhu hry som ponechala možnosť vyhľadávania v štatistikách tvrdení a rekapituláciu jednotlivých hier. Pomocou identifikátoru užívateľa, získaného zo stránky Facebook, vytváram zoznam všetkých hier hráča, s poskytnutím informácie o jazyku, čase začiatku a nahratom skóre z hry.

Metódy na vytvorenie prehľadu najlepších hráčov spomedzi užívateľových priateľov a najlepších hier obsahuje trieda *Statistics*. V oboch z metód využívam operátor *inner join*, ktorým spájam informácie, pre prípad najlepších hier, z tabuliek *games* a *users*, a v prípade najlepších hráčov z *users* a *max_score*.

6.8.2 Grafické prvky

K užívateľsky prívetivému prostrediu aplikácie prispievajú napríklad aj postupne zobrazované tipy bota alebo tiene objektov hry, vytvárajúce 3D efekt. Užívateľ má možnosť prepínať medzi jazykmi pomocou príslušných národných vlajok. V rohu hry uvádzame aj odkaz na článok o GWAP's, kde sa užívateľ môže oboznámiť s tým, ako každou svojou odohranou hrou rozširuje základňu dát určenú na vedecké ciele.

7 Návrhy do budúcnosti

Počas práce na vylepšeniach sa zriedkakedy jednalo o priamočiare riešenia bez možností výberu. Jednotlivé rozhodnutia som sa snažila robiť s ohľadom na prípadné budúce rozšírenia a v konzultácii s vedúcou svojej práce.

Jedným z takýchto prípadov bolo vytvorenie slovenskej verzie hry s dôrazom na jednoduchú rozšíriteľnosť do ďalšieho jazyka. Vzhľadom k tomu by v budúcnosti mohla byť práca rozšírená aj do anglickej verzie, a rovnako tak do svetových jazykov ako francúzština či nemčina.

Ako som spomínala v podkapitole 5.2, hra využíva bota kvôli zatiaľ relatívne malému hernému kolektívu. Pokiaľ by však hru hral dostatok ľudí, mohla by vzniknúť alternatíva pre dvoch hráčov, kde by sa hromadili CSF pri každom ťahu v hre. V tomto návrhu by bolo relevantné overovanie odlišných IP adries užívateľov, čo by vzhľadom k už existujúcemu atribútu s IP adresou nebol problém.

V budúcnosti by program mohol byť rozšírený o všeobecné skóre, ktoré by bolo súčtom jednotlivých nahratých bodov z predchádzajúcich hier. Za dosiahnutie určitých hraníc by mal užívateľ možnosť vybrať si jeden z bonusov[6], čo by ho motivovalo opätovne sa k hre vrátiť.

[6] Bonusy využiteľné neskôr v hre. Napríklad indícia navyše, zvýšený časový limit, dvojnásobné skóre z hry.

8 Záver

Cieľom tejto bakalárskej práce bolo vhodne vylepšiť hru X-plain tak, aby sa stala užívateľsky prívetivejšia, dostupná širšej verejnosti a nazbierala veľké množstvo užitočných dát.

V úvodných kapitolách práce som sa podrobnejšie venovala konceptu GWAP's, ktorý hra X-plain reprezentuje, a načrtla som vlastnosti a elementy, vedúce k vyššej kvalite hry.

Implementácii samotných vylepšení predchádzali špecifikácia a vytvorenie hernej časti programu v objektovom návrhu pomocou skriptovacích programovacích jazykov Javascript a PHP.

Výstupom je webová aplikácia prístupná v dvoch jazykových verziách, umiestnená ako mimo, tak aj na sociálnej sieti Facebook. Inštalácia Facebook PHP SDK poskytla možnosti propagácie aplikácie, konkrétne zverejnenie výsledkov hry na profile užívateľa a pozvanie priateľov do hry.

Užívateľsky atraktívnou sa hra stala hlavne prostredníctvom výsledkových tabuliek, v ktorých majú hráči možnosť vidieť porovnanie s priateľmi a súhrnné, najlepšie výsledky.

Aplikácia je odolná voči nepovolenému prístupu k databáze a k nástroju SkE, z ktorého využívame aj zabezpečené korpusy vyžadujúce autentizáciu.

Počas celého návrhu sa hra nachádzala v bezpečnom móde, v ktorom som k nej mala prístup len ja a vedúca mojej práce. Hra sa dostala na verejnosť 19.5.2013. Pred zverejnením som zaznamenala aktuálny počet nazbieraných dát, aby som za obdobie 3 týždňov zmerala účinnosť prezentovaných vylepšení. Spracované výsledky budú súčasťou prezentácie pri ústnej časti obhajoby práce.

Literatúra

[1] VON AHN, L. Games with a Purpose. *Computer*. 2006, roč. 39, č. 6, s. 92-94. DOI: 10.1109/MC.2006.196.

[2] EyeWire. [online] dostupné na URL: http://eyewire.org/ [cit. 04.05.2013], 2013.

[3] VON AHN, Luis, Mihir KEDIA a Manuel BLUM. Verbosity: a game for collecting common-sense facts. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2006, s. 75-78.

[4] NEVĚŘILOVÁ, Zuzana. X-plain - a Game that Collects Common Sense Propositions. In: *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* NLPCCS 2010. Portugal: SciTePress, 2010, s. 47-52. ISBN 978-989-8425-13-3.

[5] A slower speed of light. [online] dostupné na URL: http://gamelab.mit.edu/ games/a-slower-speed-of-light/ [cit. 19.03.2013].

[6] SMITH, Barry. Formal ontology, common sense and cognitive science. *International Journal of Human-Computer Studies*. 1995, vol. 43, 5-6, s. 641-667. DOI: 10.1006/ijhc.1995.1067.

[7] VICKREY, David, Aaron BRONZAN, William CHOI, Aman KUMAR, Jason TURNER-MAIER, Arthur WANG a Daphne KOLLER. Online word games for semantic data collection. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, s. 533-542.

[8] DETERDING, Sebastian, Dan DIXON, Rilla KHALED a Lennart NACKE. From game design elements to gamefulness: defining "gamification". In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. New York, NY, USA: ACM, 2011, s. 9-15. ISBN 978-1-4503-0816-8.

[9] ŠMERK, Pavel. K počítačové morfologické analýze češtiny [online]. Brno, 2010 [cit. 2013-05-09]. Dostupné z: https://is.muni.cz/auth/th/3880/fi_d/. Dizertačná práca. Masarykova univerzita.

[10] JAKUBÍČEK, Miloš, Vojtěch KOVÁŘ a Pavel ŠMERK. Czech Morphological Tagset Revisited. In: HORÁK a RYCHLÝ. *Proceedings of Recent Advances in Slavonic Natural Language Processing*. *Brno*: Tribun EU, 2011, s. 29-42. ISBN 978-80-263-0077-9.

[11] KILGARRIFF, Adam, Pavel RYCHLÝ, Pavel SMRŽ a David TUGWELL. The Sketch Engine. In: *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France: Universite de Bretagne-Sud, 2004, s. 105-116. ISBN 2952245703.

[12] Sketch Engine. [online] dostupné na URL: http://ske.fi.muni.cz/ bonito/run.cgi/wsketch?corpname=preloaded%2Fczes2&reload=&lemma=auto&minfreq=auto&minscore=0.0&maxitems=25&sort_ws_columns=s&clustercolls=0&structured=0&structured=1&showmwlink=0&minsim=0.15&selgrlist=a_modifier&selgrlist=gen_2&selgrlist=post_verb/ [cit. 21.04.2013].

[13] FB rastie. Má viac užívateľov i zisku, tržby ale klesli. [online] dostupné na URL: http://siit.webnoviny.sk/internet/facebook-rastie-ma-viac-uzivatelov/25717-clanok.html?from=section_article/ [cit.09.05.2013], 2006-2011.

[14] LENAT, Douglas B. CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*. 1995, roč. 38, č. 11, s. 33-38. DOI: 10.1145/219717.219745.

[15] ConceptNet5. [online] dostupné na URL: http://conceptnet5. media.mit.edu/ [cit.12.04.2013].

[16] Akinator. [online] dostupné na URL: http://en.akinator.com/ [cit.12.04.2013].

[17] Guess My Word. [online] dostupné na URL: http://www. facebook.com/ appcenter/ mysteryword/.[cit. 10.03.2013]