Četnost idiomů v textovém korpusu

Jan Bušta

Centrum zpracování přirozeného jazyka, Fakulta informatiky,
Masarykova univerzita, Brno

Úvod

Cílem práce je zjistit četnost slovesných a neslovesných hesel uvedených ve Slovníku české frazeologie a idiomatiky (Čermák – Hronek, 1994) v zadaném textovém korpusu – v tomto případě SYN2000 – a v návaznosti na zjištěné frekvence jednotlivých hesel určit poměr všech hesel v běžném textu.

Motivace

Motivací k této práci byla potřeba zjistit, které z idiomů působící problémy při počítačovém zpracování přirozeného jazyka (kvůli jejich porušování principu kompozicionality) se v textu vyskytují nejčastěji, aby bylo možné se na ně soustředit a jejich význam definovat. Nelze totiž postihnout všechna hesla, která výše zmíněný princip porušují, proto se pro další práci jeví jako nejrozumnější zpracovat právě ta nejčetnější hesla a teprve později se postupně zabývat dalšími. Dalším smyslem práce je samotné spočtení poměru idiomů vůči ostatnímu textu. Tento poměr bude zajímavým ukazatelem používání idiomatických frází v běžném textu. Aplikací informace o idiomatičnosti jazyka může být srovnávání této hodnoty s jinými jazyky. Vyhledané příklady jednotlivých frází mohou také sloužit jako přirozený, o korpus opřený podklad pro tvorbu idiomatických slovníků.

Idiomy

Před vyhledáváním hesel ze SČFI je třeba dobře rozmyslet, co vlastně v korpusu budeme hledat, zda idiomy, nebo jejich lexikální základ. V prvním případě se musíme opřít o strukturu hesla, která je v některých případech velice pevně spjata se sémantikou. Pevné spojení je např. u fixních idiomů, tj. u těch, které se vyskytují typicky pouze v jednom pádě či čísle. Idiom *ze dne na den* nelze transformovat na *ze dne na dny*. Druhou možností je tedy opomenout idiomatickou strukturu a soustředit se na lexikální obsah. Tento přístup je velice vhodný pro počítačové zpracování, na druhou stranu při hledání pouze holých slov se počet výskytů hesel (zvláště těch fixních) velice zvýší.

Přístupy k řešení

Se zvoleným přístupem jde ruku v ruce metoda kladení dotazu pomocí CQL (Rychlý, 2000). Při dotazování korpusu na jednotlivé idiomy se totiž dostáváme na rozcestí, jak vytvořit dotaz, jehož výsledkem bude chtěná četnost zadaného hesla.

Tím nejlepším řešením, které postihne každou odlišnost všech idiomů, je vytvoření speciálního dotazu pro každý idiom zvlášť. Při zvolení tohoto způsobu dotazování jsme schopni zaručit 100 % nalezení všech výskytů a zároveň nenalezení ničeho navíc. Dospějeme toho především komplexním postihnutím struktury idiomu, tzn. určíme, zda je

idiom fixní či zda je fixní jeho část, zda je možné idiom rozdělit a vložit do něj nějaký intersegment (rozvíjející přívlastek, vztažnou větu aj.), či nikoli apod. Lze vidět, že tento způsob dotazování odpovídá hledání idiomů (tedy jejich strukturních vlastností, jež jsou charakteristické pro idiomatičnost fráze). Ačkoli je tento způsob nejlepší, je v praxi nepoužitelný z důvodu, že by bylo třeba ručně připravit pro jeden každý idiom, což je časově (vzhledem k počtu hledaných hesel) nezvládnutelné a tudíž krajně nevhodné.

Dalším možným přístupem je rozdělení hesel do skupin na základě některých charakteristik. Při rozdělování idiomů do skupin je nutné přihlédnout především k možnosti vložení intersegmentu, jeho možné délce, fixnosti jednotlivých slov a jejich variabilitě – záměně jednoho slova za jiné. Při pokusu o takovéto rozčlenění ale narazíme na problém s počtem skupin, které je třeba vytvořit pro zařazení všech hesel. Ze zkušeností lze říci, že tento počet je poměrně vysoký a navíc nemusí vždy existovat jednoznačné určení skupiny. Stejně jako v předchozím případě je klíčovým předpokladem, že rozumíme všem zmíněným heslům a známe jejich možné variace. I při nejlepší snaze nelze tento předpoklad plně zajistit. Uvážíme-li, že musíme každému idiomu přiřadit nějakou skupinu, tzn. musíme projít všechna hesla slovníku, vidíme, že toto řešení je stejné jako předcházející, ale navíc s tím, že při vyhledání skupin idiomů se můžeme dopustit větších nepřesností.

Směrem, kterým se aktuálně ubírá další přístup k řešení zadaného problému, je vytvoření univerzálního dotazu, který by byl aplikovatelný na všechna hesla ze slovníku. Vytvoření takového dotazu je možné, ale výsledky budou silně spojené se zvolenou strukturou dotazu a nastavením obecných parametrů hledání. V zásadě lze říci, že dotaz je postaven na tom, zda se slova, která hledaný idiom obsahuje, nacházejí ve svém stanoveném okolí, jehož velikost je předem fixně určena. Při vyhledávání se používá nejen slovo, které idiom obsahuje, ale též jeho lemmatizovaná forma (Sedláček, 2005), tj. základní tvar (u podstatných jmen, zájmen a číslovek pokládáme za základní tvar první pád jednotného čísla, u přídavných jmen je navíc přidána podmínka prvního stupně a u sloves je lemmatizovaným tvarem infinitiv, u ostatních slovních druhů je většinou lemma totožné). Pro ilustraci: při hledání hesla *na nejvyšší úrovni* hledáme v podstatě slovo *na* (které má stejné lemma), slovo *nejvyšší* nebo všechny tvary od lemmatizovaného *vysoký* a slovo *úrovni* nebo všechny tvary od lemmatu *úroveň* vyskytující se v okolí ostatních. Toto schéma postihuje všechny výskyty hesla v korpusu (díky velké variabilitě slov a okolí pro jejich hledání), na druhou stranu přesah, který je v případě fixních idiomů značně velký, znepřesňuje (zvyšuje) četnost hledané fráze. Implementace tohoto přístupu je vzhledem k počítačovému zpracování velice efektivní. Postihnutí frází, které jsou proměnné, je velice dobré, protože dotaz dovoluje hledat slova ve proměnných vzdálenostech, což vede k lepšímu postižení intersegmentů.

Je zřejmé, že výsledek při zvoleném přístupu k hledání hesel je závislý na zvoleném dotazu, který je třeba vytvořit a nakalibrovat tak, aby rozdíl mezi výsledky při použití ručně vytvořených pravidel a jednom generalizovaném pravidlu byl co nejmenší, tedy nejpřesnější a zároveň stále dobře počítačově zpracovatelný.

Literatura:

ČERMÁK, František – HRONEK, Jiří (Eds.): Slovník české frazeologie a idiomatiky. Praha: Academia 1994.

RYCHLÝ, Pavel: Korpusové manažery a jejich efektivní implementace. [Doktorská práce.] Brno: Fakulta informatiky, Masarykova univerzita 2000.

SEDLÁČEK, Radek: Morphemic Analyser for Czech. [Doktorská práce.] Brno: Fakulta informatiky, Masarykova univerzita 2005.

Ústav Českého národního korpusu. Korpus SYN2000. [Online; 14. prosince 2008].