Filozofická fakulta Univerzity Palackého

Sestavení paralelního korpusu globálních ekonomických výhledů

(Bakalářská práce)

2019
Jakub Jílek

Filozofická fakulta Univerzity Palackého

Katedra anglistiky a amerikanistiky

Sestavení paralelního korpusu globálních ekonomických výhledů

Compilation of a parallel corpus based on global economic outlooks
(bakalářská práce)

Autor: Jakub Jílek
Studijní obor: Angličtina se zaměřením na komunitní tlumočení a překlad
Vedoucí práce: **Mgr. Michal Kubánek**

Počet stran (podle čísel): 54
Počet znaků: 72 083 (bez apendixů)
Olomouc 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a uvedl úplný seznam citované a použité literatury.

V Olomouci dne 13. srpna 2019. ..... *vlastnoruční podpis*

Děkuji vedoucímu své bakalářské práce Mgr. Michalu Kubánkovi za užitečnou metodickou pomoc a cenné rady, které mi s velkou laskavostí poskytl.

Seznam zkratek:

Zkratka	*Význam*
BNC	***British National Corpus***
BOFIT	***Institut pro transformující se ekonomiky při centrální bance Finska***
BRIC	***zkratka pro skupinu zemí***
BRL	***brazilský real***
CF	***Consensus Forecasts***
CNB	***Czech National Bank (Česká národní banka)***
CNY	***čínský renminbi***
COBUILD	***COBUILD English Advanced Learner's Dictionary (COBUILD - Slovník angličtiny pro pokročilé)***
COCA	***Corpus of Contemporary American English (Korpus současné americké angličtiny)***
ČNB	***Česká národní banka***
DBB	***Deutsche Bundesbank (Německá spolková banka)***
EIU	***Economist Intelligence Unit (Zpravodajská jednotka týdeníku The Economist)***
EK	***Evropská komise***
EU	***Evropská unie***
EURIBOR	***úroková sazba evropského mezibankovního trhu***
Fed	***Federální rezervy***
GEV	***globální ekonomický výhled***
GBP	***britská libra***
INR	***indická rupie***
JPY	***japonský jen***
LIBOR	***úroková sazba britského mezibankovního trhu***
MMF	***Mezinárodní měnový fond***

OCR	***Optical character recognition (optické rozpoznávání znaků)***
OECD	***Organizace pro hospodářskou spolupráci a rozvoj***
OSN	***Organizace spojených národů***
RUB	***ruský rubl***
TMX	***Translation Memory Exchange (formát překladové paměti)***
USD	***americký dolar***
XML	***Extensible Markup Language (rozšiřitelný značkovací jazyk)***

OBSAH

1. Úvod

Předmětem této bakalářské práce je popis kolekce textů a sestavování a využití malého specializovaného paralelního korpusu obsahujícího ekonomické výhledové zprávy, konkrétně globální ekonomické výhledy České národní banky.

Sestavováním jazykových korpusů se zabývá oblast lingvistiky zvaná korpusová lingvistika. Samotná myšlenka korpusů začala s texty v papírové podobě, poté se používaly děrné štítky, a nakonec moderní počítače. Časem se rozšířil i okruh témat textů, které jsou součástí korpusů. Z počátku byly korpusy často založeny na náboženských textech, ve 21. století lze najít korpusy různého obsahu i specializace (např. beletrie, film, terminologie aj.). Okruh uživatelů a tvůrců korpusů je dnes neomezený. Mnohé korpusy jsou dnes dostupné on-line a jsou přístupné veřejnosti.

Toto téma nabízí zajímavé spojení lingvistického výzkumu s ekonomikou.

Díky rozvoji informačních technologií se výrazně posunuly možnosti pracovat s informacemi získanými z korpusů, paralelní korpusy jsou toho příkladem. Pro účely mé bakalářské práce používám pro zarovnávání program InterText.

Cílem této práce je tudíž prozkoumat, jak probíhá proces tvorby malého paralelního korpusu a zjistit, jaké může mít mnou sestavený korpus využití. Rád bych prozkoumal možnosti využití v oblasti výzkumu funkčních stylů ve vztahu k výhledovým zprávám. Takový druh výzkumu této dosud málo prozkoumané oblasti zahrnuje popis výhledových zpráv z hlediska stylistiky, lexikologie a syntaxe.

Mým cílem je také prozkoumat možnosti extrakce textu z formátu PDF do některého z jiných formátů, který nabízí větší možnosti editace textu a který je vhodný pro tvorbu korpusů. Faktory ovlivňující extrakci jsou snadná dostupnost původních textů určených pro kompilaci korpusu, aspekt elektronické podoby, editovatelnosti, veřejné dostupnosti na internetu atd.

Globální ekonomické výhledy České národní banky jsou veřejně dostupné na internetu (každé vydání má českou a anglickou verzi, proto jsou vhodným materiálem pro paralelní korpus) a také jsem měl možnost obrátit se svými dotazy přímo na zaměstnance ***ČNB*** prostřednictvím podatelny.

Pro popis procesu sestavení paralelního korpusu bylo nejprve nutné najít informace o tom, jak je zajišťován překlad ***GEV ČNB*** do angličtiny. Kontaktoval jsem ***ČNB*** prostřednictvím e-mailu podatelny a od ***ČNB*** přišla mi relevantní odpověď spolu s kontaktem na dalšího zaměstnance.

Druhá kapitola se zabývá charakteristickými znaky korpusové lingvistiky a následně jejím možným praktickým využitím, mimo jiné v oblasti výzkumu gramatických pravidel.

Třetí kapitola se zabývá tím, jaké činnosti obnáší kompilace korpusu a možnostmi práce se specializovaným softwarem. Dále tato kapitola obsahuje popis toho, jak je možné jazykové korpusy systematicky klasifikovat. Podkapitola 3.2 je zaměřena specificky na paralelní korpusy. Tato podkapitola popisuje rozdíly mezi paralelními a srovnatelnými korpusy a obsahuje několik příkladů významných paralelních korpusů. Podkapitola 3.3 se zaměřuje na využití jazykových korpusů v oblasti translatologie. Na začátku podkapitoly je zmíněn vliv Jiřího Levého a Pražské lingvistické školy. Dále jsou zde také obsaženy teoretické koncepce M. Baker, J. Mundaye, J. Zehnalové atd. Při psaní této podkapitoly byl brán v potaz přínos jazykových korpusů i v lexikologii.

Čtvrtá kapitola se zabývá charakteristikami typu textu vybraného pro sestavení paralelního korpusu (výhledové zprávy), dále je výběr textů specifikován (***GEV ČNB*** z let 2011 a 2018) Zmiňuje i vliv systémových rozdílů mezi jazyky. Tato kapitola také popisuje, jak probíhá převod textu, vlastnosti nástroje InterText, který jsem při své práci použil, proces zarovnávání a jeho export a na závěr se stručně zmiňuje o možném využití korpusu.

Pátá kapitola popisuje možnosti využití korpusu vytvořeného v rámci této bakalářské práce vytvořil.

Zmíněn je rovněž možný přínos v oblasti odborné terminologie, výzkumu mezijazykových systémových rozdílů a stylistické klasifikace.

Tuto bakalářskou práci jsem zpracoval pro vědecké, nikoliv komerční účely. Vypracování této bakalářské práce se řídí ustanoveními zákona č. 89/2012 Sb. o vědecké licenci.

2. Korpusová lingvistika obecně

Tato kapitola se zabývá charakteristickými znaky korpusové lingvistiky a následně jejím možným praktickým využitím, mimo jiné v oblasti výzkumu gramatiky.

2.1 Korpusová lingvistika

Korpusová lingvistika patří mezi podobory lingvistiky, u kterých proběhl zásadní rozvoj ve dvacátém století s příchodem moderních počítačů. Přesněji se jedná o podobor, který se dnes zabývá analýzou kolekcí textů pomocí počítačů (ačkoliv korpusová lingvistika má kořeny i v dobách, kdy moderní počítače neexistovaly) a analyzuje tak užívaný jazyk.

Technologický vývoj ve dvacátém století nabídl nové možnosti, jak digitalizovat a elektronicky uchovávat texty, což jazykovědcům umožnilo analyzovat slova z různých jazyků novými způsoby, které v případě obyčejného zaznamenávání textů na papír nebyly možné, a zmíněný technologický pokrok navíc umožnil v počítačích skladovat větší množství jazykových dat (ať už se jedná o metadata nebo o celé miliony a miliardy slov, uložených na jednom úložišti). Na první pohled se zdá, že metody jazykového výzkumu typické pro korpusovou lingvistiku se zrodily až v době, kdy se začaly používat moderní počítače. Korpusová lingvistika je založena na empirické bázi. Je to proto, aby byly identifikovány různé lingvistické prvky a struktury, pomocí kterých lze popsat, na základě jakých pravidel funguje jednotlivý jazyk nebo více jazyků. Obecně řečeno, díky korpusové lingvistice si můžeme udělat představu, jak vypadá užívaný jazyk. Korpusová lingvistika je také užitečná, pokud chceme zkoumat frekvenci a výskyt konkrétních fonologických, lexikálních, gramatických, diskurzních a pragmatických jevů.

Nicméně, lingvistické studie prováděné s pomocí některých metod používaných v dnešních elektronických korpusech byly prováděny již před vznikem moderních počítačů, a to v oblastech jako literární studia, studia Bible, lexikografie, studia dialektů, studium jazykového vzdělávání, studiaích gramatiky atd. V případě studií Bible a beletrie můžeme uvést příklad londýnského knihkupce Alexandra Crudena, který se podílel na vydání Bible krále Jakuba. Kromě této verze Bible také vydal jako doplněk seznam konkordancí. Z oblasti lexikografie je významným počinem práce Samuela Johnsona z osmnáctého století., Tten na malé útržky papíru zaznamenával věty, které znázorňovaly užití a význam anglických slov. Samuel Johnson nashromáždil více než sto padesát tisíc citací vět pro přibližně čtyřicet tisíc heslových slov v jeho

slovníku Slovník anglického jazyka (Dictionary of the English Language). Podobným způsobem byly zpracovány podklady pro vydání Oxfordského slovníku angličtiny (Oxford English Dictionary) z roku 1928.

K zásadnímu průlomu v korpusové lingvistice došlo v padesátých letech dvacátého století. Za jednoho z průkopníků korpusové lingvistiky je považován italský kněz, teolog a filozof Roberto Busa (1913 – 2011), který pracoval téměř třicet let na vybudování korpusu díla svatého Tomáše Akvinského, italského filozofa a teologa scholastické tradice, zvaném Index Thomisticus (IBM, https://www.ibm.com, 2012).

Moderní počítače hrají v korpusové lingvistice zásadní roli. Pomocí moderních počítačů je snadnější akumulovat větší množství textů k analýze, replikovat data a riziko chybování při kompilaci korpusu se snižuje. Z historických důvodů je důležité zmínit, že korpusová lingvistika byla v minulosti prováděna „ručně“. Je to důležité zejména z hlediska rozvoje lexikografie jako oboru. Používání moderních počítačů nicméně dává lingvistům možnost rychle najít přesné informace o konkrétních slovech, frázích, úryvcích textu. Navíc dotazy lingvistů mohou být formulovány v rozhraní korpusů takovým způsobem, že je možné dotaz podrobněji specifikovat s ohledem na gramatický kontext a kolokace. Používání moderních počítačů v korpusové lingvistice vedlo k mnoha různým inovacím v oblasti lingvistického výzkumu. Nyní je snadnější provádět zpracování jazykových dat založená na matematické bázi, získaná lingvistická data jsou přesnější a kvantifikovatelnější. Díky používání moderních počítačů pro shromažďování a zpracovávání jazykových dat je výrazně jednodušší vytvářet při lingvistickém výzkumu obecné závěry. Inovace zmíněné výše mohou mít rozsáhlé praktické využití, jako např. v oblasti strojového překladu, syntézy řeči, výuky jazyků atd.

Historickými aspekty korpusové lingvistiky a proměnou trendů v čase se podrobně zabývá Graeme Kennedy ve své publikaci An Introduction To Corpus Linguistics, která byla vydána nakladatelstvím Routledge v Londýně a v New Yorku v roce 2014.

Korpusová lingvistika je charakteristická tím, že předmětem jejího zájmu jsou jazykové struktury, které je v rámci gramatických pravidel možné vytvořit, ale také se zabývá se také statistikou a pravděpodobností, respektive pravděpodobností výskytu specifických jazykových struktur.

Tvorba jazykového korpusu je velmi komplexní proces, který zahrnuje specifické aktivity. Proces, při kterém jsou texty sesbírány, uskladněny a připraveny k dalšímu užívání v rámci korpusu, se nazývá kompilace. Dále je nutné vyvinout softwarové nástroje, aby korpus mohl být řádně používán. U textů korpusů také bývá prováděn popis gramatiky a lexikonu na základě poznatků deskriptivní lingvistiky. Z československého akademického diskurzu mají pro

korpusovou lingvistiku význam především představitelé Pražského lingvistického kroužku. Ve třicátých letech dvacátého století provedli členové tohoto kroužku kvantitativní studie, které se týkaly frekvence určitých gramatických procesů, relativní frekvence různých slovních druhů, umístění a distribuceí informace ve větě a statistické distribuce typů slabik. Tyto studie byly prováděny ručně (Kennedy, 2014, 1-33.).

Díky tomu, že dnešní počítače dokážou analyzovat miliony, ba miliardy slov, lze v rámci korpusové lingvistiky) sledovat změny užívaného jazyka v průběhu času. Digitalizace a elektronizace textů ve dvacátém století umožnila jazykovědcům analyzovat jazyky novými způsoby, které u textů na papíře nebyly možné. To znamenalo pokrok jak u synchronního, tak u diachronního výzkumu jazykových jevů.

Není vhodné zaměňovat termíny „korpusová lingvistika“ a „počítačová (komputační) lingvistika“ Ačkoliv se v případě korpusové lingvistiky používají moderní počítače a specializovaný software, nejedná se v tomto případě o synonyma. Termín „počítačová lingvistika“ je historicky starší než termín „korpusová lingvistika“ a popisuje ve velké míře jevy založené na exaktních matematických principech a má obsáhlejší teoretický základ. Na druhé straně, v případě korpusové lingvistiky jsou zpracovávána data s určitou uměleckou a estetickou hodnotou jako beletrie apod. (Grishman, 1-15., 1986-).

Pod termínem jazykový korpus lze rozumět databázi textů, respektive soubor elektronicky zpracovaných a uložených dat, která slouží k primárně jazykovému výzkumu. Využitím těchto korpusů v jazykovém výzkumu se zabývá odvětví lingvistiky zvané korpusová lingvistika (Kováříková, 10., 2014).

2.2 Využití korpusové lingvistiky

Korpusová lingvistika může být využita v oblasti tvorby slovníků, rozpoznávání řeči, webových vyhledávačů apod. Souvisí s metodami zpracovávání textů a přisuzování různých vlastností jednotlivým částem textu, je užitečná rovněž při tvorbě statistik výskytu jazykových jevů. Pomocí specializovaného softwaru lze zadat specifické dotazy (Křen, 2017, http://wiki.korpus.cz/doku.php).

Korpusová lingvistika může posloužit například k účelům výzkumu gramatiky, syntaxe, standardizace jazyka, terminologie, jazykových úzů (např. jak porovnat úzy, sémantické, syntaktické, etymologické, morfologické a další rozdíly), můžeme ji využít i při výzkumech v oblasti translatologie a procesu učení jazyků. Na různých úrovních je možné zkoumat prozódii, slovní zásobu, gramatiku, druhy diskurzu nebo pragmatiku. Jazyková data v korpusech mohou zahrnovat např. data z periodik nebo beletrie,- data o mluveném jazyku (tzv. „korpusy mluveného

slova“ nebo „–mluvené korpusy“). Mluvené korpusy mohou obsahovat zvukové nahrávky nebo transkripce zvuku. Někteří autoři rozlišují korpusy a archivy textů. Korpusy jsou navrženy za účelem lingvistické analýzy a zpravidla se jedná o systematicky strukturovanou kompilaci textů. Na druhé straně, jako „archiv“ lze označit repozitář textů, ve kterém nejsou jednotlivé texty strukturovány do menších částí. (Kennedy, 70-82., 2014).

Nejmenší jednotka textu v korpusu se nazývá „token.“ Tokeny většinou mají podobu psaného slova. Někdy je jedno psané slovo rozděleno na dvě (např. mohu -li). Pro usnadnění práce s korpusy jsou obsažené texty často anotovány. Je to kvůli metainformacím o textech (původ, autor atp.) a o jazykových jevech. Příkladem anotace je lemmatizace, tj. přiřazení slovníkové podoby každému tvaru (tokenu), nebo tagování, tj. přiřazení speciální značky (tagu) popisující gramatické nebo sémantické vlastnosti slov. Slova v korpusech jsou zpravidla „tagována“ a „lemmatizována.“. Tagování a lemmatizování popisují Mareš a Pořízka.

Tagování znamená, že texty obsažené v korpusu jsou označeny identifikačními údaji (např. bibliografie nebo žánr) včetně lingvistických (např. slovní druhy, pád, číslo, rod atd.). Účelem tagování je dosáhnout desambiguace, tzn. odstranění víceznačnosti.

Lemmatizace je přiřazení různých tvarů jednoho slova k lemmatu. Termín „lemma“ označuje základní slovníkový tvar jednoho konkrétního slova (např. infinitiv, nominativ singulár apod.).

Informace, která je výstupem dotazu zadaného v korpusu, má podobu „konkordance“. Konkordance je úhrn nebo výběr výskytů daného prvku v určitém kontextu.

Jazykové korpusy jsou užitečným nástrojem při získávání dat o kolokacích. Termín „kolokace“ označuje víceslovné výrazy, zahrnuje tedy mnoho jazykových jevů – pevné vazby mezi lexikálními jednotkami (odborné termíny, víceslovné výrazy, frazémy a idiomy) a statisticky významné souvýskyty výrazů (Mareš, 18-22., 2014) (; Pořízka, 33., 71-87., 2019).

Některé korpusy svým významem převyšují jiné.

U angličtiny je významný projekt ***COBUILD*** Dictionary (viz Obr. 1). Tento elektronický slovník byl vyvinut v osmdesátých letech dvacátého století pomocí korpusu běžně mluvené angličtiny. V daném období se jednalo o revoluční projekt. První verze slovníku ***COBUILD*** byla publikována v roce 1987. Korpus, který obsahoval běžně používanou angličtinu, byl průběžně vytvářen britským vydavatelstvím Collins ve spolupráci s Birminghamskou univerzitou (University of Birmingham). Vedoucím projektu byl profesor John Sinclair. Použití korpusových dat umožnilo profesoru Sinclairovi a jeho týmu prozkoumat, jak vypadá angličtina běžně používaná lidmi, a také nabídla možnost nového způsobu strukturování hesel ve slovnících. Dobrým příkladem inovace je užití korpusových dat o frekvenci výskytu jednotlivých

slov. Tým tvůrců díky této frekvenci mohl určit, jaké významy slov jsou pro lidi učící se anglicky užitečnými (nejfrekventovanější význam bude na prvním místě). Korpus ***COBUILD*** také nově a lépe zpracovával data o anglických kolokacích, které bývaly ve slovnících zmíněny jen okrajově.

Profesor Sinclair a jeho tým rovněž rozvinul styl definicí významů slov pomocí celých vět.

Tento přístup umožňuje nejen porozumět významu daného slova, ale také vidět jeho užití v kontextu.

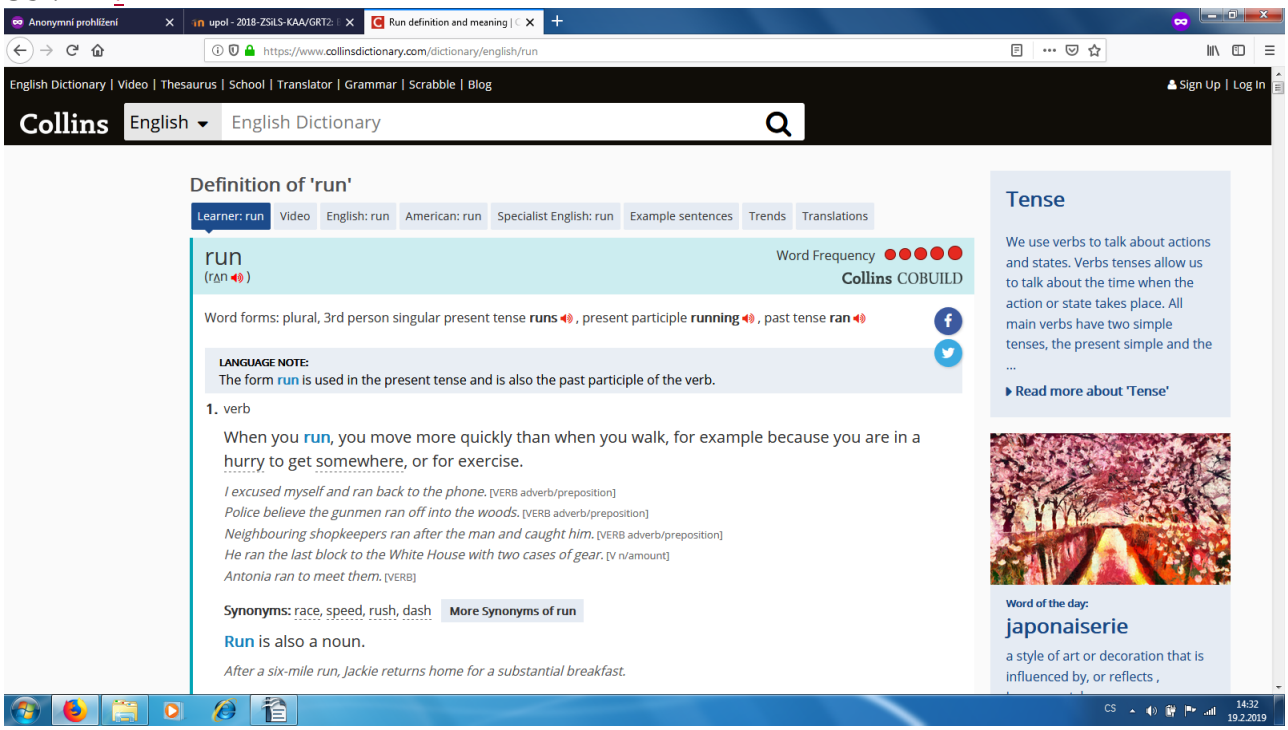
U britské angličtiny je významný i Britský národní korpus (***BNC***) (viz Obr. 2). Relevantní informace o korpusu ***BNC*** jsou dostupné na jeho webových stránkách. Jedná se o korpus mluvené a psané angličtiny, dohromady čítající přibližně sto milionů slov. Nejnovější verzí ***BNC*** je ***BNC XML*** Edition z roku 2007. Psaná část korpusu ***BNC*** (90 % obsahu) zahrnuje texty z regionálních, celonárodních i odborných článků britských periodik, akademických knih, beletrie, dopisů, memorand, školních a univerzitních esejů a dalších textů. Mluvená část (10 % obsahu) sestává z ortografických transkripcí neformálních konverzací (ty nahráli dobrovolníci různého věku, společenského postavení a z různých regionů). Jedná se o jazyk užívaný v různém kontextu, počínaje formálními obchodními a úředními setkáními a rozhlasovými pořady konče (Burnard, http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro, 2015). Korpus ***BNC*** je kódován dle směrnic sdružení TEI (Text Encoding Iniative), které reguluje standardy tagování, analýzy textů a bibliografie. Práce na korpusu ***BNC*** začala v roce 1991 a byla ukončena v roce 1994. Dosud došlo ke dvěma revizím korpusu. Další vydané edice jsou ***BNC*** World z roku 2001 a třetí edice ***BNC XML*** Edition z roku 2007. Korpus ***BNC*** má čtyři charakteristiky: Je to korpus jednojazyčný, synchronní, všeobecný a vzorkový (Burnard, http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html, 2007).

Pro americkou angličtinu je významný Corpus of Contemporary American English (***COCA***), na jehož vytvoření se významně podílel profesor lingvistiky Mark Davies z Brigham Young University v Utahu, který korpus ***COCA*** popsal na příslušných internetových stránkách (viz odkaz níže). Korpus ***COCA*** je pravděpodobně největším korpusem anglického jazyka dostupným veřejnosti. Obsahuje více než pět set šedesát milionů slov a jeho obsah je rovnoměrně rozdělen mezi mluvený jazyk, beletrii, bulvární časopisy, noviny a akademické texty. Tým z Brigham Young University také vytvořil dva nové korpusy zvané TV Corpus (325 milionů slov) a Movie Corpus (200 milionů slov). Movie Corpus má pravděpodobně ze všech korpusů vytvořených Brigham Young University nejvyšší koncentraci neformální angličtiny (Davies a kol., https://www.english-corpora.org/coca/, 2017). Dva korpusy zmíněné výše dohromady

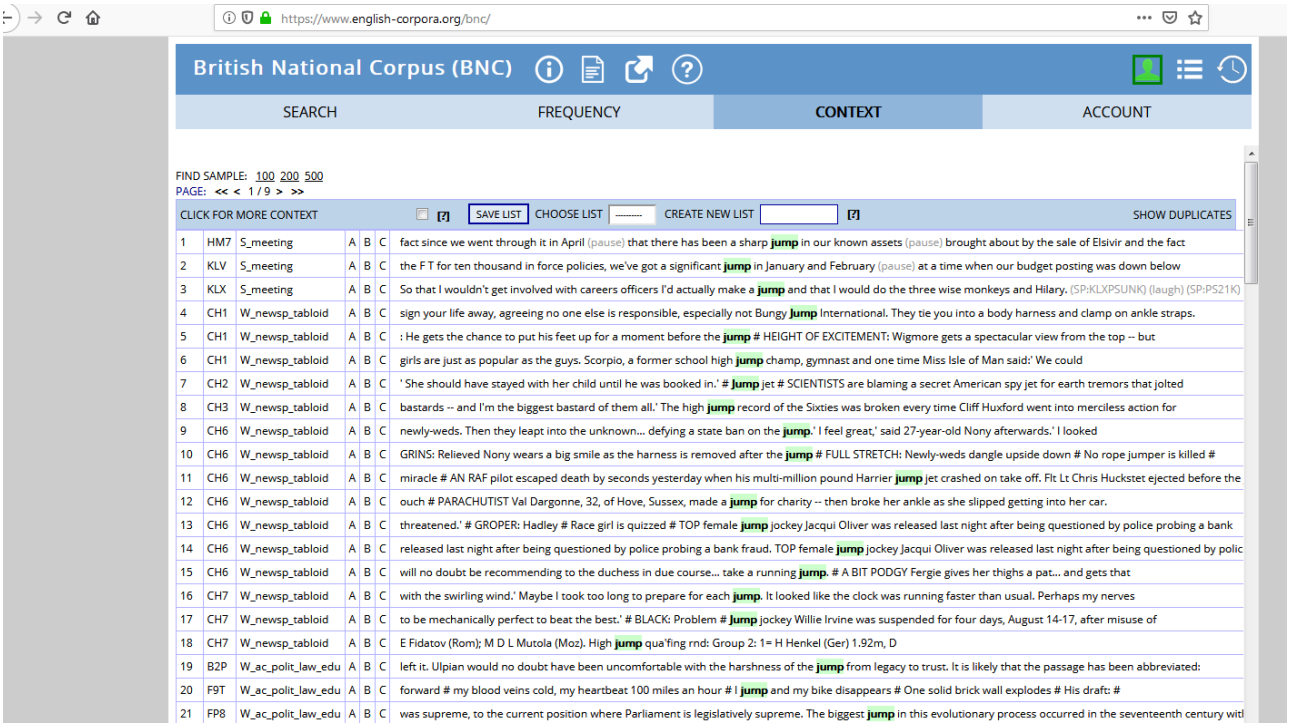
obsahují více než sto tisíc textů. Ke každému z těchto textů je přiřazen odkaz k souvisejícímu filmovému nebo televiznímu dílu na IMDB (Internet Movie Database – česky Internetová filmová databáze). Tato funkce uživatelům umožňuje vytvořit si vlastní tzv. „virtuální korpus“ pro televizní seriály a filmy. Také lze tuto funkci využít za účelem porovnání historického vývoje angličtiny a jejích dialektů. Pojem „virtuální korpus“ zastřešuje funkci korpusu, která uživateli umožní vytvořit libovolnou kolekci textů dle oblasti zájmu.

V České republice působí Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy v Praze, který spravuje projekt Český národní korpus, což je soubor jazykových korpusů (např. InterKorp, syn2015 atd., viz Obr. 3). K jeho využití je nutná registrace (zdarma přes e-mail).

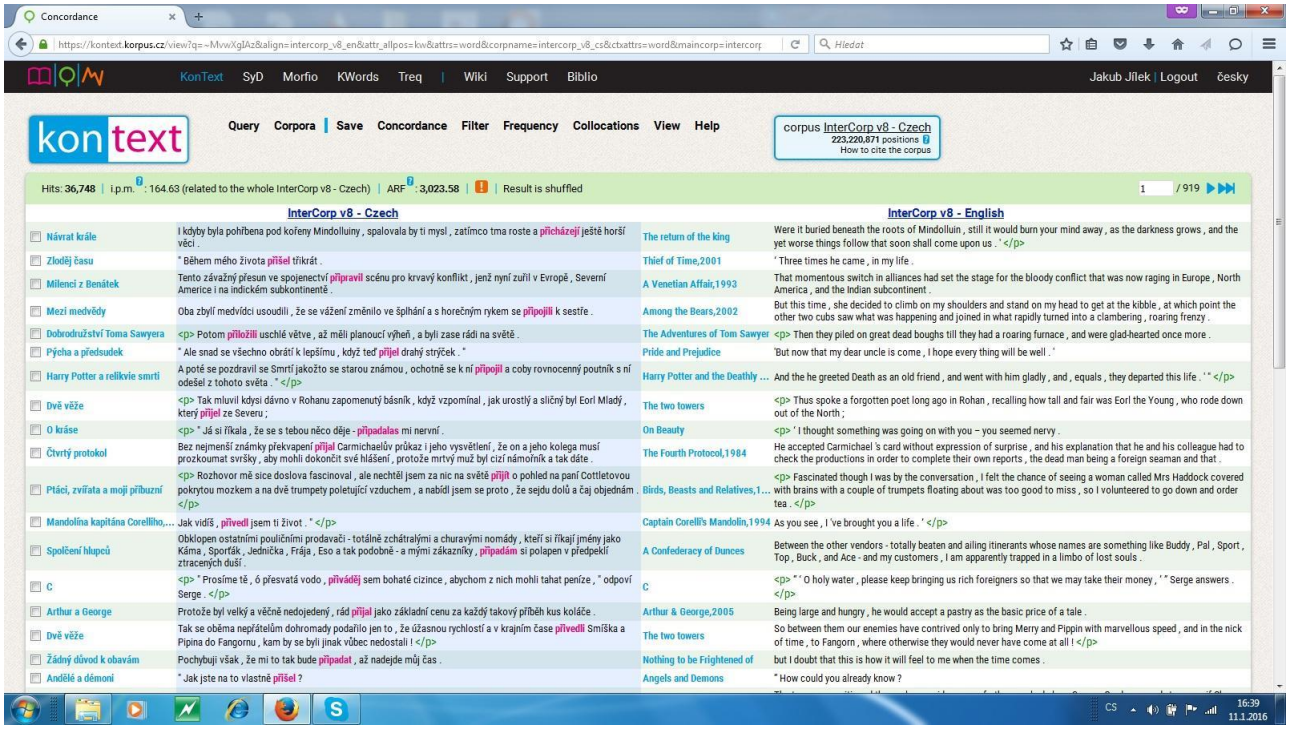
Obr. č. 1:



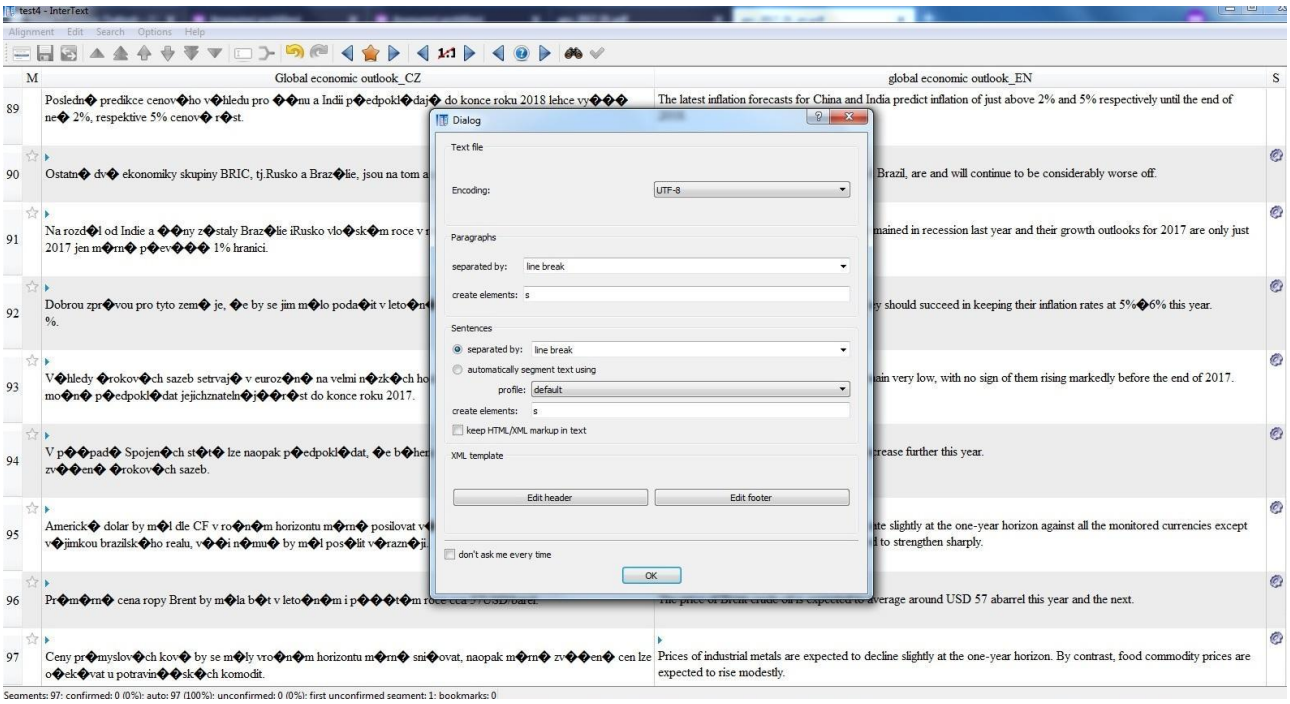
Obr. č. 2, uživatelské rozhraní BNC:


https://www.english-corpora.org/bnc/
British National Corpus (BNC)
SEARCH
FREQUENCY
CONTEXT
ACCOUNT
FIND SAMPLE: 100 200 500
PAGE: << < 1 / 9 > >>
CLICK FOR MORE CONTEXT
SAVE LIST
CHOOSE LIST
CREATE NEW LIST
SHOW DUPLICATES
1 HM7 S_meeting A B C fact since we went through it in April (pause) that there has been a sharp jump in our known assets (pause) brought about by the sale of Elsivir and the fact
2 KLV S_meeting A B C the F T for ten thousand in force policies, we've got a significant jump in January and February (pause) at a time when our budget posting was down below
3 KLX S_meeting A B C So that I wouldn't get involved with careers officers I'd actually make a jump and that I would do the three wise monkeys and Hilary. (SP:KLXPSUNK) (laugh) (SP:PS21K)
4 CH1 W_newsp_tabloid A B C sign your life away, agreeing no one else is responsible, especially not Bungy Jump International. They tie you into a body harness and clamp on ankle straps.
5 CH1 W_newsp_tabloid A B C : He gets the chance to put his feet up for a moment before the jump # HEIGHT OF EXCITEMENT: Wigmore gets a spectacular view from the top -- but
6 CH1 W_newsp_tabloid A B C girls are just as popular as the guys. Scorpio, a former school high jump champ, gymnast and one time Miss Isle of Man said:' We could
7 CH2 W_newsp_tabloid A B C ' She should have stayed with her child until he was booked in.' # Jump jet # SCIENTISTS are blaming a secret American spy jet for earth tremors that jolted
8 CH3 W_newsp_tabloid A B C bastards -- and I'm the biggest bastard of them all.' The high jump record of the Sixties was broken every time Cliff Huxford went into merciless action for
9 CH6 W_newsp_tabloid A B C newly-weds. Then they leapt into the unknown... defying a state ban on the jump.' I feel great,' said 27-year-old Nony afterwards.' I looked
10 CH6 W_newsp_tabloid A B C GRINS: Relieved Nony wears a big smile as the harness is removed after the jump # FULL STRETCH: Newly-weds dangle upside down # No rope jumper is killed #
11 CH6 W_newsp_tabloid A B C miracle # AN RAF pilot escaped death by seconds yesterday when his multi-million pound Harrier jump jet crashed on take off. Flt Lt Chris Huckstet ejected before the
12 CH6 W_newsp_tabloid A B C ouch # PARACHUTIST Val Dargonne, 32, of Hove, Sussex, made a jump for charity -- then broke her ankle as she slipped getting into her car.
13 CH6 W_newsp_tabloid A B C threatened.' # GROPER: Hadley # Race girl is quizzed # TOP female jump jockey Jacqui Oliver was released last night after being questioned by police probing a bank
14 CH6 W_newsp_tabloid A B C released last night after being questioned by police probing a bank fraud. TOP female jump jockey Jacqui Oliver was released last night after being questioned by polic
15 CH6 W_newsp_tabloid A B C will no doubt be recommending to the duchess in due course... take a running jump. # A BIT PODGY Fergie gives her thighs a pat... and gets that
16 CH7 W_newsp_tabloid A B C with the swirling wind.' Maybe I took too long to prepare for each jump. It looked like the clock was running faster than usual. Perhaps my nerves
17 CH7 W_newsp_tabloid A B C to be mechanically perfect to beat the best.' # BLACK: Problem # Jump jockey Willie Irvine was suspended for four days, August 14-17, after misuse of
18 CH7 W_newsp_tabloid A B C E Fidatov (Rom); M D L Mutola (Moz). High jump qua'fing rnd: Group 2: 1= H Henkel (Ger) 1.92m, D
19 B2P W_ac_polit_law_edu A B C left it. Ulpian would no doubt have been uncomfortable with the harshness of the jump from legacy to trust. It is likely that the passage has been abbreviated:
20 F9T W_ac_polit_law_edu A B C forward # my blood veins cold, my heartbeat 100 miles an hour # I jump and my bike disappears # One solid brick wall explodes # His draft: #
21 FP8 W_ac_polit_law_edu A B C was supreme, to the current position where Parliament is legislatively supreme. The biggest jump in this evolutionary process occurred in the seventeenth century wit

Obr. č. 3:


Concordance
KonText SyD Morfio KWords Treq Wiki Support Biblio
Jakub Jílek | Logout česky
kontext
Query Corpora Save Concordance Filter Frequency Collocations View Help
corpus InterCorp v8 - Czech
223,220,871 positions
How to cite the corpus
Hits: 36,748 | i.p.m.: 164.63 (related to the whole InterCorp v8 - Czech) | ARF: 3,023.58 | Result is shuffled
1 / 919
InterCorp v8 - Czech
InterCorp v8 - English
Návrat krále
I když byla pohřbena pod kořeny Mindolluiny , spalovala by ti mysl , zatímco tma roste a přicházejí ještě horší věci .
The return of the king
Were it buried beneath the roots of Mindolluin , still it would burn your mind away , as the darkness grows , and the yet worse things follow that soon shall come upon us . '
Zloděj času
" Během mého života přišel třikrát .
Thief of Time,2001
' Three times he came , in my life .
Milenci z Benátek
Tento závažný přesun ve spojenectví připravil scénu pro krvavý konflikt , jenž nyní zuřil v Evropě , Severní Americe i na indickém subkontinentě .
A Venetian Affair,1993
That momentous switch in alliances had set the stage for the bloody conflict that was now raging in Europe , North America , and the Indian subcontinent .
Mezi medvědy
Oba zbylí medvídci usoudili , že se vážení změnilo ve šplhání a s horečným rykem se připojili k sestře .
Among the Bears,2002
But this time , she decided to climb on my shoulders and stand on my head to get at the kibble , at which point the other two cubs saw what was happening and joined in what rapidly turned into a clambering , roaring frenzy .
Dobrodružství Toma Sawyera
 Potom přiložili uschlé větve , až měli planoucí výheň , a byli zase rádi na světě .
The Adventures of Tom Sawyer
 Then they piled on great dead boughs till they had a roaring furnace , and were glad-hearted once more .
Pýcha a předsudek
" Ale snad se všechno obrátí k lepšímu , když teď přijel drahý strýček . "
Pride and Prejudice
'But now that my dear uncle is come , I hope every thing will be well . '
Harry Potter a relikvie smrti
A poté se pozdravil se Smrtí jakožto se starou známou , ochotně se k ní připojil a coby rovnocenný poutník s ní odešel z tohoto světa . "
Harry Potter and the Deathly ...
And the he greeted Death as an old friend , and went with him gladly , and , equals , they departed this life . ' "
Dvě věže
 Tak mluvil kdysi dávno v Rohanu zapomenutý básník , když vzpomínal , jak urostlý a sličný byl Eorl Mladý , který přijel ze Severu ;
The two towers
 Thus spoke a forgotten poet long ago in Rohan , recalling how tall and fair was Eorl the Young , who rode down out of the North ;
O kráse
 " Já si říkala , že se s tebou něco děje - připadalas mi nervní .
On Beauty
 ' I thought something was going on with you – you seemed nervy .
Čtvrtý protokol
Bez nejmenší známky překvapení přijal Carmichaelův průkaz i jeho vysvětlení , že on a jeho kolega musí prozkoumat svršky , aby mohli dokončit své hlášení , protože mrtvý muž byl cizí námořník a tak dále .
The Fourth Protocol,1984
He accepted Carmichael 's card without expression of surprise , and his explanation that he and his colleague had to check the productions in order to complete their own reports , the dead man being a foreign seaman and that .
Ptáci, zvířata a moji příbuzní
 Rozhovor mě sice doslova fascinoval , ale nechtěl jsem za nic na světě přijít o pohled na paní Cottletovou pokrytou mozkem a na dvě trumpety poletující vzduchem , a nabídl jsem se proto , že sejdu dolů a čaj objednám .
Birds, Beasts and Relatives,1...
 Fascinated though I was by the conversation , I felt the chance of seeing a woman called Mrs Haddock covered with brains with a couple of trumpets floating about was too good to miss , so I volunteered to go down and order tea .
Mandolína kapitána Corelliho,...
Jak vidíš , přivedl jsem ti život . "
Captain Corelli's Mandolin,1994
As you see , I 've brought you a life . '
Společenství hlupců
Obklopen ostatními pouličními prodavači - totálně zchátralými a churavými nomády , kteří si říkají jmény jako Káma , Sporťák , Jednička , Frája , Eso a tak podobně - a mými zákazníky , připadám si polapen v předpeklí ztracených duší .
A Confederacy of Dunces
Between the other vendors - totally beaten and ailing itinerants whose names are something like Buddy , Pal , Sport , Top , Buck , and Ace - and my customers , I am apparently trapped in a limbo of lost souls .
C
 " Prosíme tě , ó přesvatá vodo , přiváděj sem bohaté cizince , abychom z nich mohli tahat peníze , " odpoví Serge .
C
 " ' O holy water , please keep bringing us rich foreigners so that we may take their money , ' " Serge answers .
Arthur a George
Protože byl velký a věčně nedojedený , rád přijal jako základní cenu za každý takový příběh kus koláče .
Arthur & George,2005
Being large and hungry , he would accept a pastry as the basic price of a tale .
Dvě věže
Tak se oběma nepřátelům dohromady podařilo jen to , že úžasnou rychlostí a v krajním čase přivedli Smíška a Pipina do Fangornu , kam by se byli jinak vůbec nedostali !
The two towers
So between them our enemies have contrived only to bring Merry and Pippin with marvellous speed , and in the nick of time , to Fangorn , where otherwise they would never have come at all !
Žádný důvod k obavám
Pochybuji však , že mi to tak bude připadat , až nadejde můj čas .
Nothing to be Frightened of
but I doubt that this is how it will feel to me when the time comes .
Andělé a démoni
" Jak jste na to vlastně přišel ?
Angels and Demons
" How could you already know ?

Obr. č. 4:


test4 - InterText
Alignment Edit Search Options Help
M
Global economic outlook_CZ
global economic outlook_EN
S
89
Posledn predikce cenov ho v hledu pro Indii p edpokl daj do konce roku 2018 lehce vy
ne 2%, respektive 5% cenov r st.
The latest inflation forecasts for China and India predict inflation of just above 2% and 5% respectively until the end of
90
Ostatn dv ekonomiky skupiny BRIC, tj.Rusko a Braz lie, jsou na tom a
Brazil, are and will continue to be considerably worse off.
91
Na rozd l od Indie a ny z staly Braz lie iRusko vlo sk m roce v r
2017 jen m rn p ev 1% hranici.
mained in recession last year and their growth outlooks for 2017 are only just
92
Dobrou zpr vou pro tyto zem je, e by se jim m lo poda it v leto n
%.
y should succeed in keeping their inflation rates at 5% 6% this year.
93
V hledy rokov ch sazeb setrvaj v euroz n na velmi n zk ch ho
mo n p edpokl dat jejichznateln j r st do konce roku 2017.
ain very low, with no sign of them rising markedly before the end of 2017.
94
V p pad Spojen ch st t lze naopak p edpokl dat, e b her
zv en rokov ch sazeb.
rease further this year.
95
Americk dolar by m l dle CF v ro n m horizontu m rn posilovat v
v jmkou brazilsk ho realu, v i n mu by m l pos lit v razn ji.
te slightly at the one-year horizon against all the monitored currencies except
l to strengthen sharply.
96
Pr m rn cena ropy Brent by m la b t v leto n m i p t m
erage around USD 57 abarrel this year and the next.
97
Ceny pr myslov ch kov by se m ly vro n m horizontu m rn sni ovat, naopak m rn zv en cen lze
o ek vat u potravin sk ch komodit.
Prices of industrial metals are expected to decline slightly at the one-year horizon. By contrast, food commodity prices are
expected to rise modestly.
Segments: 97: confirmed: 0 (0%); auto: 97 (100%); unconfirmed: 0 (0%); first unconfirmed segment: 1; bookmarks: 0
Dialog
Text file
Encoding:
UTF-8
Paragraphs
separated by:
line break
create elements:
s
Sentences
separated by:
line break
automatically segment text using
profile:
default
create elements:
s
keep HTML/XML markup in text
XML template
Edit header
Edit footer
don't ask me every time
OK

3. Korpusy

Tato kapitola se zabývá tím, jaké činnosti obnáší kompilace korpusu, a specializovaným softwarem. Dále tato kapitola nastiňuje možnosti systematické klasifikace jazykových korpusů. Podkapitola 3.2 je zaměřena na paralelní korpusy. Tato podkapitola popisuje rozdíly mezi paralelními a srovnatelnými korpusy a obsahuje několik příkladů významných paralelních korpusů. Podkapitola 3.3 se zaměřuje na využití jazykových korpusů v translatologii. Ze začátku je zmíněn vliv Jiřího Levého a Pražské lingvistické školy. Dále obsahuje teorie M. Baker, J. Mundaye, J.Zehnalové atd. Při psaní této podkapitoly jsem vzal v potaz přínos jazykových korpusů i v lexikologii.

3.1 Kompilace a dělení korpusů

Co se týče autorských práv, je v některých případech nutné požádat vlastníky práv k původním textům o svolení k jejich použití, zpravidla při kompilaci korpusu pro komerční účely. Jak uvádí Evans, mnoho korpusů vytvořených pro komerční účely obsahuje rozmanité druhy textů a získání práv na používání textů může být finančně a časově náročné. Jiná situace je, pokud někdo vytváří korpus v malém rozsahu a pro vlastní potřebu. Pokud žádáme vlastníky autorských práv k povolení použít texty v korpusech, lze zmínit, že: a) texty budou použity pouze za účelem výzkumu, nikoli za účelem dosažení zisku, b) je limitovaný okruh osob, který bude mít přístup k danému korpusu, c) po dokončení výzkumu budou korpusová data vymazána, d) nebudou používány kompletní texty, ale pouze některé z jejich částí. Také je důležité konzultovat náklady na autorská práva s tím, kdo výzkum financuje (např. ústavy, nadace, univerzity atd.). Dle Evanse mohou veškeré argumenty uvedené výše usnadnit proces získání autorských práv k použití originálních textů (Evans, 2007, 1-5).

Základním kamenem kompilace jazykových korpusů je nahrát texty, které bude nově vytvořený korpus obsahovat, do počítače v elektronické podobě. Vzhledem k tomu, jak fungují softwarové programy určené~~ý~~ pro tvorbu korpusů, je vhodným formátem textu pro korpusy formát TXT, nikoliv jiné formáty jako DOCX nebo PDF. Mezi programy používané pro kompilaci korpusových dat patří InterText, Sketch Engine (dostupný online) aj.

Při kompilaci korpusu je důležité si uvědomit, že nepracujeme s originály textů, ale s kopiemi. Ke kompilaci korpusu jsou používány i texty, které před kompilací nebyly dostupné v elektronické podobě. Pokud chceme takový text použít, můžeme ho naskenovat a získat textová data pomocí ***OCR*** (optické rozpoznávání znaků) softwaru. Určitou roli hraje i vzhled a formátování stránek skenovaného textu (text psaný tučně nebo kurzívou, font písma atd.). Mnoho skenovaných dokumentů obsahuje netextové prvky, což zahrnuje grafy, diagramy, obrázky apod.

Při kompilaci korpusu je třeba vzít v úvahu, jaké informace o vzhledu stránky a formátu textu a netextových prvků chceme zachovat.

Je vhodné položit si otázku, v čem spočívá přínos zmíněného specializovaného softwaru. Uživatel má možnost provést u textů zarovnání (angl. „alignment“). Zarovnání znamená, že zdrojový a cílový text se rozdělí na segmenty a ideálně jsou všechny sousedící segmenty zdrojového a cílového textu zarovnány tak, aby dva segmenty vedle sebe byly ekvivalenty ve významu. V této bakalářské práci je pro zarovnání použit program InterText, který rozděluje texty na segmenty. Nedostatky v automatickém zarovnání jsou následně odstraněny ručně. Také existuje možnost přiřadit k textům relevantní metadata nebo anotace.

Při extrahování textu za účelem kompilace existuje riziko, že dojde ke ztrátě nebo dekontextualizaci určitých informací. Pokud taková situace nastane, je možné tyto nedostatky kompenzovat přidáním anotace. Před přidáním anotace je nutné provést lemmatizaci lexémů (viz lemmatizace ve druhé kapitole).

Pomocí anotace lze přidat navíc relevantní informace pro zefektivnění práce s korpusem. To, jaký charakter mají jednotlivé anotace, je podmíněno strukturou jednotlivých textů. Text v korpusu se skládá ze dvou základních částí – první je „záhlaví“ (angl. header) a druhou je „tělo textu“ (angl. body). Záhlaví stránky zpravidla obsahuje metadata, která zahrnují informace jako jméno autora, název díla, osobní informace o autorovi a mluvčích v textu atd. Přínosem metadat je možnost poskytnout extralingvistické informace. Další druhem anotace, který se vyskytuje v těle textu je tagování (viz tagování ve druhé kapitole). Proces tagování spočívá v tom, že každé slovo uvnitř textu je označkováno specifickým kódem, který dodává konkrétní informace. Typickým příkladem tagování je tagování slovních druhů označováno zkratkou „POS“ („part of speech“ z angličtiny). Dalším příkladem tagování je dělení textu na odstavce pomocí špičatých závorek na odstavce: <p> <\p> a na věty: <s> <\s>. Pro tagování se používá značkovací jazyk ***XML*** (Extensible Markup Language), který byl vyvinut a standardizován konsorciem W3C (World Wide Web Consortium). Návrh tohoto jazyka vychází ze staršího standardu SGML (Standard Generalized Markup Language). Z tohoto standardu také vychází formát dokumentů HTML (Hyper-Text Markup Language). Jak dále zmiňuje Richta, značky mají tvar obecných závorek (např. otevírací/zavírací...) (Richta, 2008, 15.).

Korpusy lze klasifikovat podle různých kritérií. Korpusy lze rozlišit na paralelní a srovnatelné. Paralelní korpus je korpus, který obsahuje původní, zdrojové texty v jazyce A a jejich překlady v jazyce B, případně ve více jazycích. Paralelní korpus může být jednosměrný (uni-directional), tj. může obsahovat pouze překlady z jazyka A do jazyka B, nebo obousměrný (bi-directional), tj. může obsahovat překlady z jazyka A do jazyka B i překlady z jazyka B do jazyka A.

Srovnatelný korpus se skládá z částí (subkorpusů), které byly sestaveny dle stejných kritérií výběru textů/vzorků a jsou obdobně vyvážené a reprezentativní. Co se týče vícejazyčného srovnatelného korpusu, ten zahrnuje originální, původně psané texty ve více jazycích (ne zdrojové texty a jejich překlady). Jednojazyčný srovnatelný korpus lze rozdělit na dvě části: překladovou a nepřekladovou. Obsahuje tedy dva subkorpusy, jeden s texty původně psanými, nepřekladovými (non-translated) a druhý s texty překladovými (translated) (Chlumská, 2014, 221-232.). V kontextu historického vývoje jazyku lze rozdělit korpusy synchronní a diachronní. Synchronní korpus usiluje o záznam jazykového úzu v jednom (většinou v relativně úzce vymezeném období). Na rozdíl od korpusů diachronních tedy neposkytuje možnosti zkoumání jazykového vývoje, příp. jen ve velmi omezené míře. Z pohledu současného jazyka se jako synchronní jeví korpus, který zachycuje jazyk živý, tj. takový, který je užíván žijícími mluvčími. Na rozdíl od korpusu synchronního, který je vždy vytvářen pro určité omezené období, se diachronní korpus snaží zachycovat proměny jazykového úzu v čase. Z pohledu registru, žánru a témat textů lze rozdělit korpusy na všeobecné a specializované (např. právo, Korán, Evropský parlament apod.). Za „vzorkový“ korpus lze považovat korpus, který se skládá ze vzorků textů, jejichž délka zpravidla nepřekročí čtyřicet pět tisíc slov.

Finální podoba korpusu závisí také na tom, jaký má účel a z jakého okruhu textů se skládá.

Kvalitu korpusu může ovlivnit mnoho faktorů. Nicméně, při kompilaci korpusu jsou zásadní především tři faktory: velikost korpusu, vyváženost korpusu a jeho reprezentativnost. Velikost korpusu má vliv na to, jaká data můžeme získat. Obecně platí, že čím větší je korpus, tím lepší má využití. Pokud je ale zkoumán jazykový jev, který má vysokou frekvenci výskytu, nemusí nutně být malá velikost korpusu problém.

Pokud chceme vytvořit korpus beletristických překladů za účelem výzkumu používaného jazyka, je třeba vzít v úvahu faktory jako rok vydání a vyhotovení překladu, žánr literárního díla, forma vyprávění, z jaké části nebo kapitoly literárního díla text pochází atd. Mezi vybranými texty musí být správný poměr vlastností, aby mohly být získány relevantní informace. Během takového časového rozpětí se jazyk mohl značně proměnit. Proto je nutná vyváženost korpusu.

Reprezentativnost korpusu je odvozena od toho, do jaké míry se z něho dají vyvodit pravidla všeobecně aplikovatelná. Tzn., že pokud se v korpusu objevuje nějaké gramatické pravidlo, nastává otázka, zda je dané pravidlo aplikovatelné na jazyk jako takový. S reprezentativností také souvisí termín „saturace textu v lexikální rovině“. Termín saturace textu vyjadřuje situaci, kdy je jeden text rozdělen na několik částí o stejném počtu slov. Do jaké míry je text saturovaný

lze určit podle toho, kolik nových lexikálních jednotek se vyskytuje v přidané části textu o stejném počtu slov. Počet lexikálních jednotek by měl být v každé z částí rozděleného textu přibližně stejný.

V případě korpusů mluveného slova je velkou výhodou, že během posledních tří desetiletí došlo k výraznému posunu v oblasti digitalizace zařízení, která slouží k nahrávání a zaznamenávání zvuku. Nicméně, pokud je cílem sesbírat data o jazyku používaným v mluvené komunikaci, která by měla být autentická, může mít chování mluvčích vliv na výsledek (mluvčí se nemusí vyjadřovat pro ně standardním způsobem, pokud jsou si vědomi toho, že jejich verbální projevy jsou nahrávány a jsou předmětem výzkumu~~?~~). Předmětem výzkumu mé bakalářské práce je kompilace malého paralelního korpusu, žánr použitých textů je ekonomický výhled (Chlumská, 2014, 221-232.).

3.2 Paralelní korpusy

Paralelní korpusy mají velké využití v kontrastivní lingvistice. Zdrojové texty a jejich kvalitní překlady lze považovat za ukázku vyrovnaného dvojjazyčného výstupu (balanced bilingual output), protože data o cílovém a zdrojovém jazyce se zásadně neliší od dat, která by poskytly texty vytvořené rodilými mluvčími daných jazyků. Vezmeme-li v úvahu strukturalistické přístupy, kde je funkce gramatických kategorií definována na základě vztahu vůči jiným kategoriím, tak překladové ekvivalenty mají navzájem korespondující slovní druh a je vhodné je takto srovnávat. Existuje totiž nebezpečí, že lingvista bude srovnávat dva velmi rozdílné nebo nesouvisející jazykové jevy (Martinková, 2014, 270-285.). Mezi významné paralelní korpusy patří v českém prostředí InterCorp, který je součástí projektu Český národní korpus, významný je také projekt OPUS2.

3.3 Využití korpusu při překladatelském procesu a při překladatelském výzkumu

Teoretické poznatky o vlivu technologického pokroku na proces překladu publikovaly i významné osobnosti české translatologie. Zehnalová ve svém příspěvku uvádí Jiřího Levého, pokračovatele tradice Pražské lingvistické školy. Jiří Levý během svého života naznačil, jak se bude translatologie vyvíjet. V rámci své publikační činnosti propojoval translatologii s kybernetikou, informatikou a teorií her. Těžištěm Levého odborného zájmu byl styl textu a převod stylu v překladu.

Korpusy se začaly používat v translatologii v devadesátých letech. V roce 1993 Mona Baker publikovala článek, který přinesl zásadní zlom v pohledu na využití korpusů – článek

„Corpus Linguistics in Translation Studies: Implication and Applications“. Mona Baker uvádí, že za účelem výzkumu ekvivalence v překladu je nutné zkoumat nejen konkrétní zdrojový a cílový text, ale skupinu textů určitého typu, aby bylo možné udělat závěr o stylistické a funkční ekvivalenci. Pokud chceme v cílovém textu reprodukovat nejen formální jazykové struktury, ale také dát textu myšlenku a určit prioritu obsažených informací, je nutné studovat autentické texty, které mají stejný diskurz jako překládaný text (Baker, 1993, 233-237.). V této bakalářské práci jsou proto kromě výhledových zpráv ***ČNB*** zmíněny i výhledové zprávy jiných organizací. Podle Mundaye je dnes korpusová lingvistika součástí translatologických studií. Munday uvádí, že je problém najít metodologii výzkumu, která by byla všeobecně akceptována. To je částečně způsobeno tím, že zvolená metodologie nevyhnutelně závisí na předmětu výzkumu, a také tím, že dnes se používají korpusy i pro jiné účely než lexikografické, pro které byly původně zamýšleny (např. tvorba internetových překladačů) (Munday, 2016, 231.).

Saldanha a O'Brien uvádí, že první etapa translatologických výzkumů se zaměřovala zejména na rozdíly mezi překlady a zdrojovými texty. Druhá etapa výzkumů nabízí mnoho nových možností, okruh výzkumných otázek je daleko širší, specializovanější, a to se také projevuje při výzkumu stylistiky (Saldanha a O'Brien, 2014, 1-9.). Laviosa zmiňuje, že zpravidla jsou při výzkumech používány korpusy s velkým objemem dat, ale za určitých okolností je přínosné použít „malé“ korpusy, např. pro výzkum beletrie. Termín „malý“ korpus v tomto kontextu znamená, že se jedná o korpus vytvořený za účelem použití při jednom specifickém výzkumu. Další charakteristikou malých korpusů je, že se skládají z textů konkrétního žánru, autora, literárního období, periodika, instituce apod. Malé paralelní korpusy mohou být využity jako překladové paměti (obvykle ve formátu ***TMX***) a být exportovány do překladových studií (CAT nástrojů). Zehnalová a Munday se zaměřují na použití paralelních korpusů v překladech beletrie. Překladová studia však nejsou pro překlad beletrie doporučovány, vhodnější jsou pro technické nebo ekonomické překlady. Překladové paměti mohou být použity i opačně – překladové paměti ve formátu ***TMX*** je možné použít při sestavování nových malých korpusů, např. v programu InterText nebo Sketch Engine, který Zehnalová řadí mezi korpusové manažery. Korpusové manažery mají jiné funkce než CAT nástroje. Korpusové manažery obsahují funkce tagování, funkce rozpoznávání slovních druhů a nástroje lemmatizace. Díky možnosti pracovat s více jazyky jsou paralelní korpusy užitečnými nástroji v oblasti kontrastivní lingvistiky.

Jedná se o podobor lingvistiky, u něhož se o produkci teoretických poznatků významně zasloužila Pražská škola zmíněná výše. Johansson, který je jedním z tvůrců Anglicko-norského paralelního korpusu (ENPC), zmiňuje ve svém článku Seeing through multilingual corpora vliv této školy na vytváření paralelních korpusů (Zehnalová, 2018, 101.-105.). Lingvisté vidí přínos korpusů v tom, že překladatelé mají lepší přístup k překladatelským řešením jiných

překladatelů, paralelní zarovnání textů umožňuje porovnat stylistiku, překladatelé si mohou v paralelním korpusu vybrat z více překladatelských řešení. Překladatelé běžně pracují s více slovníky. Pokud však slovníky nenabízí ekvivalent v cílovém jazyce, mohou paralelní korpusy posloužit jako zdroj informací. Pokud překladatel nenajde ekvivalent v cílovém jazyce, má možnost vytvořit neologismy, což může mít negativní vliv na kvalitu překladu. Překladová paměť je de facto specifický typ paralelního korpusu, ve kterém jsou zdrojový a cílový text zarovnány tak, aby spolu jednotlivé segmenty z hlediska významu korespondovaly. Způsob, jakým jsou překladové paměti přístupné v uživatelském rozhraní, se liší mezi jednotlivými překladovými studii. Ale překladová studia nenabízí typy funkcí běžně dostupné u korpusů (např. vyhledat konkordance termínu a kontext, ve kterém se obvykle používá). V dnešní době většina překladových studií obsahuje konkordanční funkci, která překladateli umožňuje najít jednotlivé příklady toho, jak byla určitá lexikální jednotka, větná konstrukce nebo fráze přeložena.

Paralelní korpusy mají využití v oblasti lexikografie. Pokud je paralelní dvojjazyčný korpus použit při tvorbě slovníku, poskytuje lexikografovi informace, které se týkají užití slov v určitém kontextu. Tudíž, pokud je hledán správný překladatelský ekvivalent pro termín v cílovém jazyce a existuje více různých variant, paralelní korpusy mohou pomoci zvolit tu správnou variantu. V oblasti bilingvní lexikografie jsou paralelní korpusy velkým přínosem. Proces tvorby slovníku začíná výběrem slovníkového hesla v cizím jazyce. Poté je dané heslo přeloženo do cílového jazyka. V případě více možných ekvivalentů je dobrou strategií dát ten nejvhodnější, respektive nejfrekventovanější překlad hesla na první místo. Třetí krok při tvorbě slovníku se nazývá syntéza. Syntéza znamená, že do slovníku je vložen záznam hesla a jeho překladu a tento záznam se ve slovníku nachází na specifické pozici (např. seřazení hesel podle abecedy, podle tematických okruhů slovní zásoby atd.) (Héja, 2012, 1-3.). Velkou roli při tvorbě slovníků hraje také intuice lexikografů. Jak uvádí Héja, pokud mají lexikografové problém nalézt ten správný ekvivalent v cílovém jazyce slovníku, spoléhají na svou intuici, což může způsobit problémy.

Přínos paralelních korpusů v oblasti lexikografie spočívá v tom, že díky korpusům nemusí lexikografové tolik spoléhat na svou intuici. Další výhodou je, že lexikální jednotky, které jsou předmětem zpracování výzkumu, mohou být v paralelních korpusech přehledně zobrazeny ve mnoha různých kontextech. Jaký je správný překladový ekvivalent pro heslo je možné vyhodnotit i díky tomu, že paralelní korpusy poskytují informace o frekvenci a pravděpodobnosti výskytu lexikálních jednotek v určitém kontextu (Héja, 2012, 1-3.).

Paralelní korpusy lze využít i pro výzkum syntaxe a následně překladatelských strategií jako je explicitace a jiné (Olohan, 2002, 153-169.).

Podle Marca a Van Lawickové mají korpusy pro překladatele dvojí využití: a) jako nástroj dokumentace a b) jako prostředek pro tvorbu výukových materiálů. V oblasti výuky se v podstatě jedná o vylepšení pracovního postupu, který používají překladatelé odborných textů už léta používání referenčních textů při řešení terminologických problémů. Překladatelé mohou používat texty v papírové verzi, ale díky korpusům se snadno a rychle dostanou ke většímu množství informací, takže pokud se studenti translatologie naučí pracovat s korpusem, je to pro ně velmi užitečné (Marco a van Lawick, 2009, 10.). Paralelní korpusy jsou užitečné při výuce subžánrů jazyka, např. právního (Scott, 2012, 87-99.).

Frankenberg-Garcia a Santos uvedli do provozu portugalsko-anglický paralelní korpus Compara. Compara má kódování a možnosti zarovnání textu, která uživatelům umožňují vidět poznámky překladatelů a zjistit informace o tom, kdy a kde se překladatel rozhodl spojit, rozdělit, vymazat nebo přeskládat věty. Malé paralelní korpusy mají využití při zkoumání kulturně-specifických prvků. Compara je tzv. otevřený (open-ended) korpus – objem textu průběžně narůstá. Původně se ale jednalo o korpus zaměřený jen na beletrii a kulturně-specifické prvky (Frankenberg-Garcia a Santos, 2003,71-72.).

Pearsonová vytvořila takový korpus, který obsahuje populárně-naučné vědecké články a jejich překlady. Podle Pearsonové to ukazuje, že paralelní korpusy hrají důležitou roli v odborných znalostech překladatelů. Také se domnívá, že paralelní korpusy jsou komplementární vůči srovnatelným korpusům. (Pearson, 2003, 15-24.)

4. Metoda práce

Tato kapitola se zabývá charakteristikami typu textu vybraného pro sestavení paralelního korpusu (výhledové zprávy), dále je výběr textů specifikován (**GEV ČNB** z let 2011 a 2018). Tato kapitola také popisuje, jak probíhá převod textu, vlastnosti nástroje InterText, proces zarovnávání a jeho export.

4.1 Charakteristika výhledových zpráv

Tato podkapitola charakterizuje výhledové zprávy. Obsahuje informace o výhledových zprávách mezinárodních organizací, bank, nadnárodních firem a podrobně popisuje globální ekonomické výhledy České národní banky. U ***GEV*** popisuje strukturování textu, okruh obsažených témat a rozdíly mezi verzemi z roku 2011 a z roku 2018.

Výhledové zprávy jsou druhem textu, jehož hlavním účelem je informovat o budoucím vývoji v určité oblasti. Mezi vydavatele výhledových zpráv patří: státní instituce (např. ministerstva), finanční a bankovní instituce (***ČNB***, JP Morgan), mezinárodní organizace (např. ***OECD, MMF***, ***EU***...) a soukromé společnosti (např. Deloitte, Moody's). Obsah výhledů finančních a bankovních institucí, vychází z role dané instituce na trhu.

Česká národní banka je tzv. centrální banka. Centrální banka je zřízena jako státní instituce, která má regulativní funkci v oblasti měnové politiky oficiální měny své země, stanovuje výši úrokových sazeb a provádí dozor nad komerčními bankami a dalšími subjekty finančního trhu v rámci dvojúrovňového bankovního systému. Komerční banky jsou banky, obvykle v soukromém vlastnictví, které poskytují finanční služby ostatním subjektům na trhu (firmy, domácnosti apod.). Příkladem takové banky je Československá obchodní banka (Dittrichová a Ptatscheková, 41-42., 2013).

Organizace pro hospodářskou spolupráci a rozvoj (***OECD***) se ve svých výhledech zabývá budoucím vývojem světové ekonomiky, stejně tak jako Světová banka a Mezinárodní měnový fond.

Ekonomické výhledy ***OECD*** mají podobu prezentací v programu MS PowerPoint, doplněných handoutem. Na úvodním slidu prezentace bývá uveden název, datum a autor, v případě ***OECD*** INTERIM ECONOMIC OUTLOOK je to hlavní ekonom ***OECD***. Následující druhý slide obsahuje názvy klíčových témat, na které se výhled zaměřuje (např. Growth is weakening particularly in Europe, Vulnerabilities in China, Europe and financial markets could derail the global economy apod). Informace ve výhledech jsou znázorněny sloupcovými diagramy, spojnicovými grafy atd. Na konci prezentace jsou popsána klíčová témata v bodech. Další informace jsou v handoutu.

Jedny z nejobsáhlejších ekonomických výhledových zpráv vydává Světová banka. V rámci čtyř kapitol zprávy najdeme i jednu esejisticky zaměřenou, ale nejvíce vybočuje velmi obsáhlá kapitola, která popisuje světovou ekonomiku podle makroregionů (např. Subsaharská Afrika, Střední a Jižní Amerika, Střední východ atd.) (World Bank Group, I-XI., 2019).

Mezinárodní měnový fond vydává jednou měsíčně výhledy světové ekonomiky, které mají nekonzistentní strukturu, ale první kapitola má pravidelný název Global Prospects and Policies (International Monetary Fund, 1., 2019). Ačkoliv se jedná o organizaci dlouhodobě etablovanou na mezinárodní scéně, ***ČNB*** opakovaně kritizuje přesnost jejich předpovědí.

Evropská komise se ve svých výhledových zprávách zabývá analýzou členských států. Na výhledových zprávách EK s názvem European Economic Forecast se podílí Generální ředitelství pro hospodářské a finanční záležitosti. Výhled začíná úvodníkem, pro každý členský stát je obsažen odborný článek o jeho ekonomické situaci a na konci výhledu jsou seznamy grafů a tabulek (***EU***, 1-29. 2019.).

Deloitte je nadnárodní korporace poskytující rozsáhlé služby pro firmy, hlavně auditorské. Tato společnost vydává ekonomické výhledy o světové ekonomice i o ekonomice jednotlivých zemí. Ekonomické výhledy jednotlivých zemí včetně České republiky mají tuto strukturu: a) úvodní slovo, b) komentář ke globální ekonomické situaci, c) komentář k vybrané zemi (ČR), který obsahuje vývoj ekonomických indikátorů, tzn. míra nezaměstnanosti, inflace, HDP, d) shrnutí, e) data v tabulkách. Naproti tomu, výhledy specializované na kapitálové trhy jsou bohatší co do obsahu a rozsahu. Struktura pro rok 2019 je následující:

a) Retail banking

b) Corporate banking

c) Transaction banking

d) Investment banking

e) Payments

f) Wealth management payments

g) Market infrastructure

Každá kapitola má svůj podnadpis (např. Corporate banking: Digitalisation and a new credit discipline). Tematické okruhy výhledů jsou vybrány tak, aby je bylo možné provázat se službami společnosti Deloitte.

Společnost Moody's je ratingová agentura, která hodnotí, nakolik perspektivní jsou pro investování a obchodování státy a jiné tržní subjekty. Na internetových stránkách Moody's je popsána metodologie ratingu. Ta zahrnuje práci s logaritmy, matematickými konstantami, proměnnými, koeficienty, rozdíly mezi přirozeným logaritmem atd. (Deloitte, 1-12.; 1-26.).

Další centrální bankou, která vydává ekonomické výhledy, je americká centrální banka Federal Reserve, známá pod zkratkou ***Fed***. Je zajímavé, že dvě centrální banky mohou mít odlišnou koncepci ekonomických výhledů. Ekonomické výhledy ***Fedu*** mají podobu internetové stránky, na které je zaznamenán projev některého z bankovních funkcionářů.

V případě ekonomického výhledu pro duben 2019 se jedná o místopředsedu Richarda Claridu (v ***ČNB*** tato funkce nejspíš odpovídá funkci viceguvernéra). Původní médium, jakým byl ekonomický výhled sdělen, bylo tedy mluvené slovo, text projevu byl následně na stránkách ***Fedu*** zveřejněn s datem a místem projevu. Název tohoto ekonomického výhledu je „U.S. Economic Outlook and Monetary Policy" (Výhled americké ekonomiky a monetární politiky). Projev byl přednesen na konferenci Washington Policy Summit. Začátek projevu místopředseda banky zahájil poděkováním organizátorům. Následoval popis ekonomické situace USA (konjunktura nebo recese?), poté se řečník přesunul k ekonomickým vyhlídkám světové ekonomiky a zmínil možná rizika (např. Brexit). Další částí projevu je měnová politika. Ohledně měnové politiky řečník zmínil kroky, které americká centrální banka nedávno podnikla. Poskytnuté informace se týkají schůzí Výboru banky (obdobným orgánem v ***ČNB*** je zřejmě Bankovní rada). Tato část projevu také obsahuje ekonomické indikátory a názor řečníka, jaký mohou mít na ekonomiku vliv.

Také jsou zde popsány plány banky do budoucna (např. koncepce rozvah banky). Celý výhled končí poděkováním publiku. Projev obsahuje odkazy a poznámky. Tyto citace obsahují: a) poznámky, že některé z příspěvků řečníka jsou osobní názory, které nutně nemusí reflektovat pozici ***Fedu***, b) odkazy na jiné zdroje, c) informace o tom, co ***Fed*** plánuje v budoucnu zveřejnit.

Česká národní banka se ve svých ***GEV*** zabývá světovou ekonomikou. Věnuje se ekonomickým výhledům podle světových regionů, výhledům kurzů měn, komoditám, indikátorům trhu, výši úrokových sazeb, výši HDP, inflaci a předpovědím mezinárodních institucí. Každý výhled má speciální kapitolu zaměřenou na vybrané téma. Některé výhledy jsou doplněny jednou kapitolou zaměřenou na aktuální téma. ***GEV ČNB*** vychází jednou měsíčně.

Česká národní banka má jiný přístup než ***Fed***. Na internetových stránkách ***ČNB*** je specializovaná sekce, kde jsou zveřejňovány výlučně globální ekonomické výhledy ve formátu PDF, každý výhled má zvlášť českou i anglickou verzi. V této sekci si lze stáhnout výhledy aktuální i výhledy z minulosti (nejstarší globální ekonomický výhled ke stažení je z ledna 2011). Globální ekonomické výhledy ***ČNB*** vycházejí pravidelně jednou měsíčně. Pracují na něm vybraní experti, mnozí dlouhodobě a s vybranou doménou specializace a rozdělenými funkcemi v rámci pracovního týmu. Vydáváním ***GEV*** se zabývají sekce měnová a statistiky a odbor vnějších ekonomických vztahů ***ČNB***. Výhledy ***ČNB*** mají pevně danou strukturu.

Kapitola „SHRNUTÍ“ obsahuje informace o důležitých ekonomických ukazatelích a předpovědi jejich budoucího vývoje. Tyto ukazatele zahrnují HDP, inflaci, předstihové ukazatele, sazby mezibankovního trhu (***EURIBOR***, ***LIBOR*** atd.), výhled kurzů měn a výhled cen komodit.

Kapitola PŘEDPOVĚDI MEZINÁRODNÍCH INSTITUCÍ se dělí na podkapitoly: II.1 HDP, II.2 Porovnání předpovědi HDP a změna oproti předchozí předpovědi, II.3 Inflace a II.4 Porovnání předpovědi inflace a změna oproti předchozí předpovědi. Ve všech podkapitolách se předpovědi týkají těchto čtyř regionů: Eurozóna, USA, Německo a Čína.

Podkapitola II.1 obsahuje předpovědi růstu HDP vybraných teritorií. Předpovědi jsou doplněny spojnicovými grafy a tabulkami.

Podkapitola II.2 obsahuje výpověď o tom, pro jaké regiony se předpovědi zhoršily nebo zlepšily. Jsou zde obsaženy i data v podobě tabulek a sloupcových diagramů.

Podkapitola II.3 obsahuje předpovědi o inflaci, které jsou doplněny spojnicovými grafy a tabulkami.

Podkapitola II.4 obsahuje informace o tom, jaké jsou nové výhledy růstu inflace ve vybraných teritoriích. Jsou zde také obsaženy i data v podobě tabulek a sloupcových diagramů.

Ve druhé kapitole jsou jako zdroje použity výpočty ***ČNB*** s použitím databáze Eurostat, ***CF***, ***MMF, OECD, EK, ECB, Fed, DBB a BOFIT***.

Třetí kapitola obsahuje informace o předstihových ukazatelích (indikátorech), respektive o vývoji jejich růstu nebo poklesu. Analyzované regiony opět zahrnují eurozónu, USA, Německo a Čínu. Jako zdroje ***ČNB*** používá své výpočty s použitím databáze ***OECD***, dále daty z ***EK***, institutu IFO a University of Michigan.

Kapitola o výhledu úrokových sazeb se dělí na dvě podkapitoly: IV.1 Výhled krátkodobých a dlouhodobých úrokových sazeb: eurozóna a IV.2 Výhled krátkodobých a dlouhodobých úrokových sazeb: USA. První podkapitola se zabývá prognózami úrokových sazeb ***EURIBOR 3M, EURIBOR 1R*** a také výnosem německého vládního dluhopisu Bund 10R.

Druhá kapitola je zaměřena na úrokové sazby ***USD LIBOR 3M, USD LIBOR 1R*** a výnos amerického vládního dluhopisu Treasury 10R. Tabulky a spojnicové grafy jsou obsaženy v obou podkapitolách. Kurzy vybraných měn, kterým se tato kapitola věnuje, jsou tyto: a) americký dolar vůči euru, b) japonský jen vůči americkému dolaru, c) americký dolar vůči britské libře, d) švýcarský frank vůči americkému dolaru.

Kapitola s názvem „Výhled cen komodit“ se dělí na dvě podkapitoly, z nichž je jedna specificky zaměřena na ropu a zemní plyn, čili dva specifické druhy komodit. Klasifikace komodit je taktéž doplněna tabulkou a spojnicovým grafem, který znázorňuje tři hodnoty výhledu cen ostatních komodit, a to sice 1) výhledu cen průmyslových kovů, 2) výhledu cen potravinářských komodit a 3) výhledu cen komoditního koše celkem.

Témata kapitoly „Zaostřeno na…“ se pokaždé proměňují a také se u této kapitoly nejčastěji střídají autoři. Ve srovnání s ostatními kapitolami je více publicistického rázu, ostatní jsou spíše rázu zpravodajského, zprostředkovávají čtenáři věcná fakta o ekonomickém vývoji na trhu, a to s pomocí přesných algoritmů, které se opakují napříč jednotlivými vydáními globálních ekonomických výhledů České národní banky. Všechny kapitoly kromě „Zaostřeno na...“ jsou konzistentní i z hlediska okruhu témat. U ***GEV ČNB*** se jedná o texty psané odborným stylem, také označovaném jako styl vědecký nebo naučný. Jedná se o texty praktického rázu, neboť obsažené informace využívají profesionálové v oblasti ekonomie a financí (Minářová,141., 2011, 2013.). Na úplném konci globálních ekonomických výhledů ***ČNB*** se nachází seznam zkratek. Redakční tým globálních ekonomických výhledů ***ČNB*** sestává z odborníků se specializacemi v oblastech kurzů měn, obchodování s komoditami, analýz derivátů finančního trhu, finanční stability, měnové politiky, rozpočtové ekonomiky, ekonomiky pracovního trhu, makroekonomie atd. Časem došlo u globálních ekonomických výhledů České národní banky ke změně koncepce, až dospěly do aktuální podoby v roce 2018.

Popis verzí ***GEV ČNB*** z roku 2018:

Druhá kapitola se zabývá ekonomickým výhledem ve vyspělých zemích, tzn. země eurozóny, Německo zvlášť, Spojené státy, Spojené království a Japonsko. V případě eurozóny patří mezi zkoumané indikátory růst HDP, inflace, předstihové ukazatele a úrokové sazby. Hodnoty znázorňují spojnicové grafy a sloupcové diagramy. V případě jednotlivých vyspělých zemí patří mezi zkoumané indikátory růst HDP a inflace, obojí znázorněno spojnicovými grafy. Zdroje informací jsou: ***CF, MMF, OECD, DBB, Fed, EIU***, Bank of England a Bank of Japan. Třetí kapitola se zaměřuje na ekonomický růst v zemích skupiny ***BRIC***. Tato skupina zemí je zpracovaná stejně jako vyspělé země v předcházející druhé kapitole. Zdroje informací jsou: ***CF***, ***MMF***, ***OECD*** a ***EIU***. Čtvrtá kapitola prošla největší proměnou. Předstihové ukazatele (Markit, Nikkei atd.) a výhledy kurzů měn (***USD, GBP, JPY, CNY, INR, RUB, BRL***) jsou popsány pouze spojnicovými grafy a tabulkami. Struktura páté a šesté kapitoly je stejná. ***GEV ČNB*** mají přílohy označené písmeny A1 až A5 (seznam zkratek se stal součástí příloh).

4.2 Role systémových rozdílů mezi jazyky

Mimo extralingvistických odlišností se vyskytuje specifikum jazykové – v anglických překladech se častěji vyskytují frázová slovesa.

Sestavený česko-anglický paralelní korpus lze využít při výzkumu systémových rozdílů mezi češtinou a angličtinou. Takovýchto rozdílů je celá řada. Z hlediska typologie je čeština syntetický (převládají jednoslovné gramatické tvary) a angličtina analytický jazyk (převládají analytické gramatické tvary, resp. gramatické kategorie vyjadřovány pomocnými

slovy) (Chauncy Fowler, 1850, 20-22.). Dalším autorem, který se zabývá systémovými rozdíly mezi oběma jazyky je Dagmar Knittlová v publikaci K teorii i praxi překladu, vydané v roce 2000. Knittlová uvádí následující:

V důsledku analytičnosti a syntetičnosti existují– mezi češtinou a angličtinou četné rozdíly. Angličtina má větší tendence k víceslovným gramatickým tvarům a víceslovným pojmenováním, čeština má zase větší tendence k jednoslovným gramatickým– tvarům a jednoslovným pojmenováním. Existují i rozdíly ve skloňování podstatných jmen, slovosledu a časování sloves.

Ve vyjadřování informací má angličtina tendence k nominálnosti (vyjadřování pomocí podstatných a přídavných jmen) a čeština k verbálnosti (slovesné vyjadřování, tzn. pomocí plnovýznamových sloves v určitém tvaru). Dobrým příkladem nominálnosti angličtiny jsou tzv. sponová slovesa, tedy sémanticky chudá slovesa doplněna jménem.

Co se týče rodů, v angličtině převládá pojetí přirozeného rodu a nižší stupeň gramatikalizace. V češtině je ale rod plně rozvinutá gramatická kategorie a typickým jevem české gramatiky je shoda podmětu s přísudkem.

Velké rozdíly najdeme také v syntaktické rovině. V angličtině se častěji vyskytují osobní zájmena, protože je zde nutné vyjádřit podmět, zatímco v češtině existuje i podmět nevyjádřený. Problémem při překladu z angličtiny do češtiny bývá překlad časů. Angličtina má bohatou soustavu slovesných časů, v češtině jsou na výběr tři časy. Pokud čeština příslušným časem nedisponuje, dochází často k nesprávnému překladu nebo zanedbání kompenzace (hl. předminulého času) např. časovým adverbiem. Proto je ekonomický překlad výzvou pro překladatele (Knittlová, 2000, 33-81.).

4.3 Okruh textů pro sestavení korpusů

Podkladem pro sestavení česko-anglického paralelního korpusu výhledových zpráv ve byly globální ekonomické výhledy České národní banky z let 2011 a 2018, stažené formátu PDF.[1] V předkládané bakalářské práci jsou uvedeny také příklady jiných výhledových zpráv, které výsledný paralelní korpus neobsahuje. Výhledové zprávy ***ČNB*** byly zvoleny, protože výchozím jazykem je autentická čeština od rodilých mluvčích, a kvůli obsaženým ekonomickým okruhům. Co se týče slovní zásoby, u ***GEV ČNB*** vidím velký přínos v seznamech zkratek, protože české a anglické zkratky pro stejný jev nejsou vždy identické (např. b barel = bbl barrel, b. b. bazický bod (setina procentního bodu) = bp basis point (one hundredth of a percentage point) atd). Tento seznam zkratek v některých případech obsahuje i vysvětlení českého významu anglické zkratky.

Jak bylo uvedeno v předchozích kapitolách, globální ekonomické výhledy České národní banky slouží k předpovědi ekonomického vývoje, jsou tudíž využívány při makroekonomickém plánování vlády a centrální banky. Vzhledem k dlouhodobé tradici a fungování ***ČNB*** jsou překlady ***GEV ČNB*** do angličtiny velmi kvalitní a jsou dobrým zdrojem informací. I z těchto důvodů byly zvoleny tyto výhledové zprávy. Každý ze zdrojových souborů má specifický název. Systém pojmenování dokumentů je popsán v tabulce č. 1. Tabulky č. 2 a č. 3 obsahují data a statistiky o souborech ve formátu TXT, nahraných do InterTextu.

Tabulka č. 1:

gev_	YYYY_	MM	(_en)
ZAČÁTEK NÁZVU	**ROK**	**MĚSÍC**	**PŘÍPONA PRO ANGLICKÉ PŘEKLADY**

Tabulka č. 2:

Tabulka č. 2 a) – čeština ve formátu TXT:

[1] Odkazy: https://www.cnb.cz/cs/menova_politika/gev/, https://www.cnb.cz/en/monetary_policy/geo/. Datum stažení: 9.3.2019 – GEV 2011 a GEV 12/2018; 24.11.2018 – GEV 01-11/2018.

MĚSÍC	ROK	JAZYK	VELIKOST V KB	POČET SLOV
leden	2011	CZ	26	3373
únor	2011	CZ	32	4187
březen	2011	CZ	31	4017
duben	2011	CZ	36	4442
květen	2011	CZ	42	5001
červen	2011	CZ	33	3946
červenec	2011	CZ	36	4293
srpen	2011	CZ	50	5883
září	2011	CZ	46	5471
říjen	2011	CZ	36	4247
listopad	2011	CZ	33	4061
prosinec	2011	CZ	33	3921
leden	2018	CZ	44	5802
únor	2018	CZ	42	5569
březen	2018	CZ	55	7334
duben	2018	CZ	50	6605
květen	2018	CZ	55	7207
červen	2018	CZ	51	6691
červenec	2018	CZ	53	6813
srpen	2018	CZ	48	6090
září	2018	CZ	60	8448
říjen	2018	CZ	42	5744
listopad	2018	CZ	63	8365
prosinec	2018	CZ	52	6265

Tabulka č. 2 b) – angličtina ve formátu TXT:

MĚSÍC	ROK	JAZYK	VELIKOST V KB	POČET SLOV
leden	2011	EN	27	3835
únor	2011	EN	33	7587
březen	2011	EN	32	4514
duben	2011	EN	33	4697
květen	2011	EN	39	5529
červen	2011	EN	31	4388
červenec	2011	EN	35	4988
srpen	2011	EN	47	6645
září	2011	EN	43	6149
říjen	2011	EN	34	4727
listopad	2011	EN	31	4646
prosinec	2011	EN	33	4536
leden	2018	EN	46	7206
únor	2018	EN	45	6978
březen	2018	EN	50	8054
duben	2018	EN	45	7129
květen	2018	EN	50	7971
červen	2018	EN	45	7136
červenec	2018	EN	49	7504
srpen	2018	EN	44	6659
září	2018	EN	55	8949
říjen	2018	EN	39	6228
listopad	2018	EN	56	8901
prosinec	2018	EN	41	6493

Tabulka č. 3:

CELKOVÁ VELIKOST KORPUSU	KB	SLOV
	2032	285224
	KB	SLOV
VELIKOST KORPUSU – ROK 2011_CZ	434	52842
VELIKOST KORPUSU – ROK 2011_EN	418	62241
VELIKOST KORPUSU – ROK 2018_CZ	615	80933
VELIKOST KORPUSU – ROK 2018_EN	565	82002
	KB	SLOV
PRŮMĚRNÁ VELIKOST KORPUSU – ROK 2011_C	45,67	5800,33
PRŮMĚRNÁ VELIKOST KORPUSU – ROK 2011_E	15,67	2225,25
PRŮMĚRNÁ VELIKOST KORPUSU – ROK 2018_C	27,50	5870,83
PRŮMĚRNÁ VELIKOST KORPUSU – ROK 2018_E	259,92	35906,50

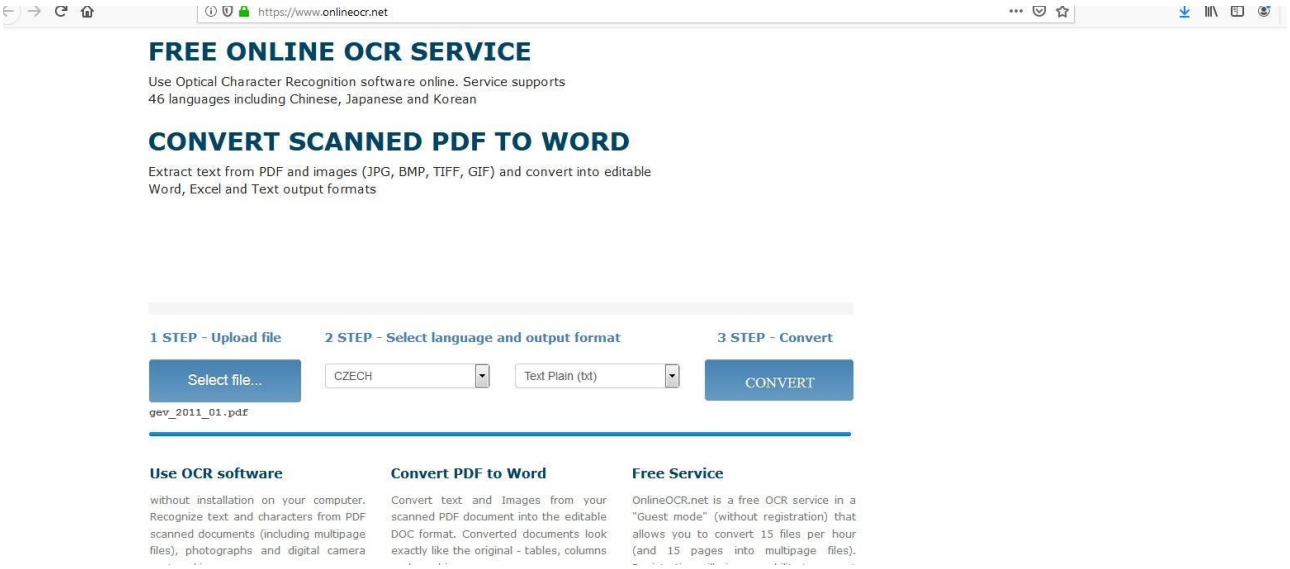
Obr. č. 5, ***GEV*** na webu stránkách ***ČNB***:

4.4 Převod textu

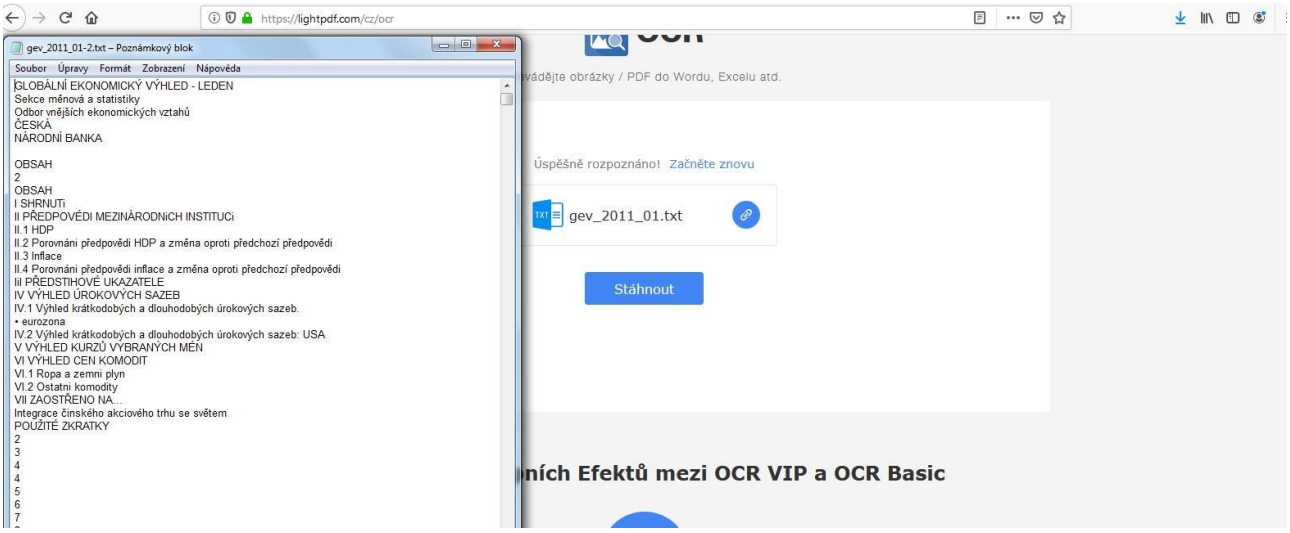
Aby bylo možné texty nahrát do příslušného programu, je nutné provést určité úpravy. Na internetu je zdarma dostupných mnoho ***OCR*** programů, které lze použít při převodu.

Při výběru online ***OCR*** nástroje je třeba dát si pozor, jestli ***OCR*** funguje v jazycích, které uživatel potřebuje. Také je důležitý formát, v němž se mají rozpoznané znaky vyexportovat. V rámci této bakalářské práce byl zvolen formát TXT, aby texty mohly být nahrány do programu InterText. Použito bylo vícero online nástrojů. Jak probíhal proces ***OCR***, ukazují následující obrázky:

Obr č. 6, nahrání dokumentu:



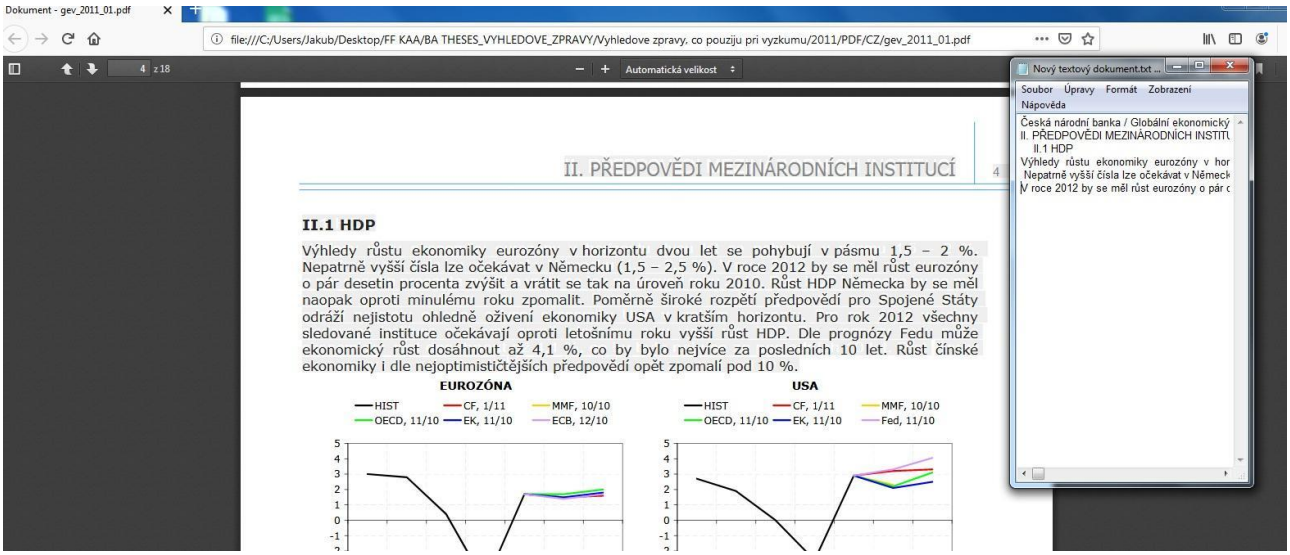
Obr č. 7:



Po první fázi, při níž byly použity zmíněné ***OCR*** nástroje, bylo nutné změnit pracovní postup. Vyexportované soubory ve formátu TXT obsahovaly řadu nepřesností, jako chybné rozpoznání malých a velkých písmen, nežádoucí změny formátování v číslování stran ***GEV ČNB***

a v datech z obsažených grafů. Vzhledem k úspoře času a snaze extrahovat informace co nejefektivněji, byly nadále relevantní úseky textu kopírovány z formátu PDF do formátu TXT ručně. PDF dokumenty jsem si otevřel v internetovém prohlížeči a relevantní úseky textu jsem vykopíroval do souboru ve formátu TXT. Z originálních textů byly zkopírovány: úvodní strana, obsah, informace o týmu zpracovatelů, texty kapitol a podkapitol s jejich nadpisy, poznámky pod čarou a seznamy zkratek. Zkopírovány však nebyly seznamy článků uvedených na konci ekonomických výhledů a až na některé výjimky také úseky textu, které se vztahovaly k obsaženým grafům.

Obr č. 8, ruční extrahování textu:



Zdroj: „FREE Online OCR - Convert PDF to Word or Image to text.“ Online OCR, zobrazeno 10.3.2019. https://www.onlineocr.net/

Pokud se u extrakce textu z formátu PDF není vyextrahovaný text rozdělen správně mezerami a je ho třeba upravit, aby korespondoval s rozvržením stránky, nadpisy a odstavci výchozího textu, je nutné rozdělit segmenty textu tak klávesou Enter, která následující segment textu přesune na nový řádek. Při zarovnávání textu ve specializovaném programu jako InterText, je taková úprava nutná.

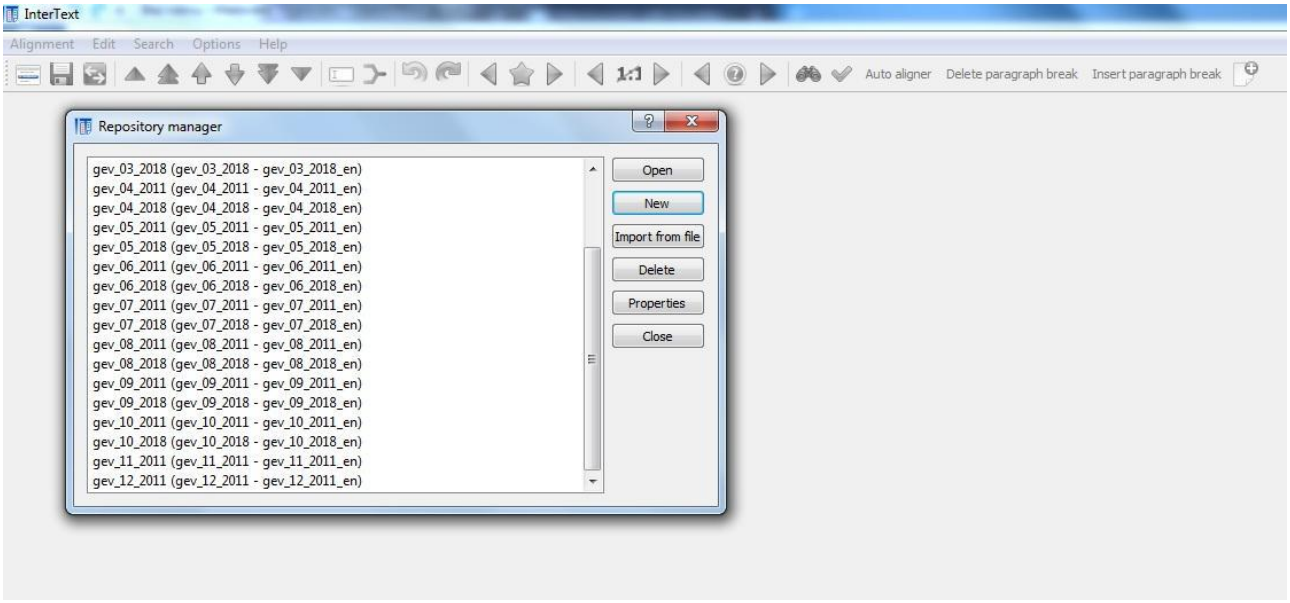
4.5 InterText

InterText je program volně dostupný, lze ho nalézt zdarma ke stažení na serveru https://wanthalf.saga.cz/. Jedná se o osobní stránky Pavla Vondřičky, českého lingvisty z Ústavu Českého národního korpusu. InterText byl vytvořen za účelem tvorby paralelních korpusů, které mají potenciální využití při překladu. Tento program byl vyvinut pro práci na projektu InterCorp.

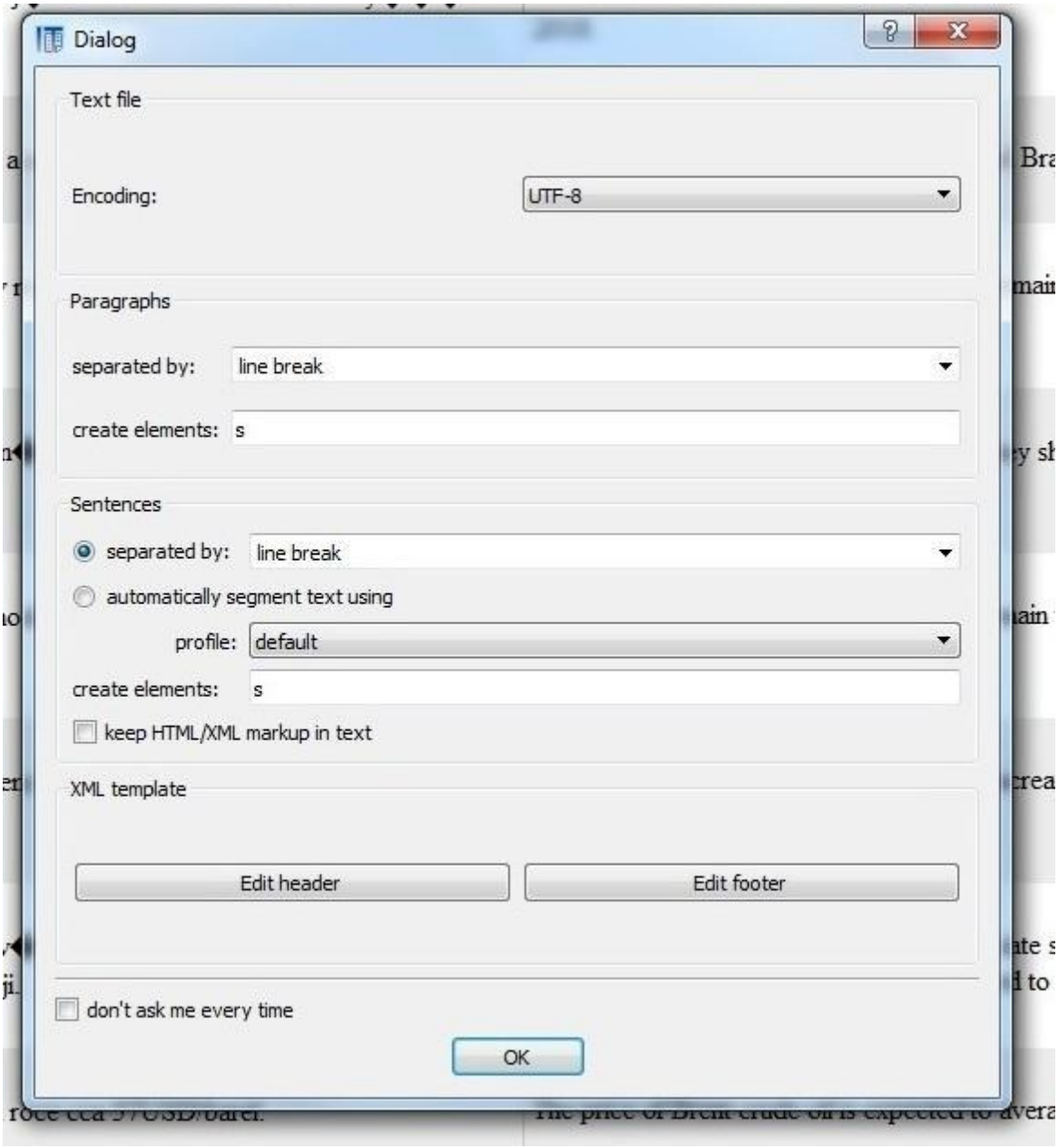
Kromě existujících souborů je možné provádět zarovnání i u prázdných dokumentů, které mohou být editovány (v nastavení je nutné zvolit možnost „empty document“ místo „local file“), následně se v programu zobrazí dva prázdné sloupce, do kterých lze vkládat text a elementy. Co se týče formátu souborů, InterText může být použit pro práci s dokumenty ve formátu ***XML*** a v případě nahraných TXT souborů pro zarovnání je nejvhodnějším kódováním UTF-8. Pokud je u nahraných souborů nastaven jiný typ kódování, je velmi pravděpodobné, že formátování segmentů bude—hodně nepřesné a některé z obsažených znaků budou rozpoznány chybně nebo vůbec (místo znaku se objeví otazník).

Součástí tohoto programu je repozitář obsahující všechna zarovnání. Repozitář se otvírá ikonou v levém horním rohu obrazovky. Repozitář kromě zobrazení existujících zarovnání uživatelům umožňuje importovat zarovnání ze souboru, viz obrázek č. 9. Aby byly paralelní texty připraveny k zarovnání v co nejvyšší kvalitě, je nutné udělat určitá nastavení v dialogovém oknu, jak ukazuje obrázek č. 9.

Obr. č. 9, repozitář v InterTextu:



Obr. č. 10, nastavení InterTextu v dialogovém oknu:



Kódování UTF-8 je třeba nastavit v dialogovém okně. Dále je nutné zvolit způsob oddělení odstavců (prázdným řádkem nebo koncem řádků, jak to je v mém případě). Jako profil je nastaven profil výchozí (default). Políčko „keep HTML/***XML*** markup in the text“ se TXT dokumentů netýká – pokud je při nahrávání zaškrtnuto, soubor nepůjde nahrát a program nahlásí chybu v parsování segmentů.

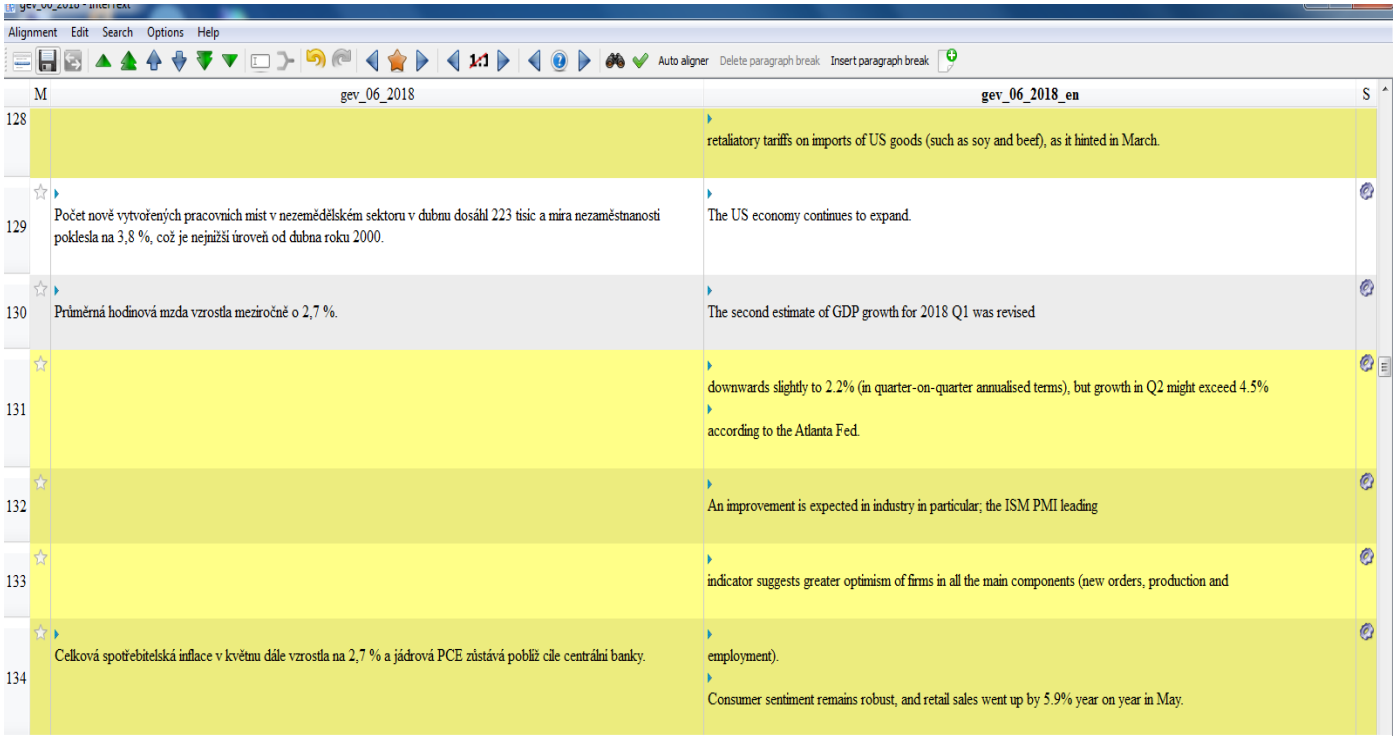
4.6 Proces zarovnávání a jeho export

Jak může vypadat soubor s paralelními texty na začátku procesu zarovnávání, ukazují následující obrázky č. 11, 12 a 13:

Obrázek č. 11:

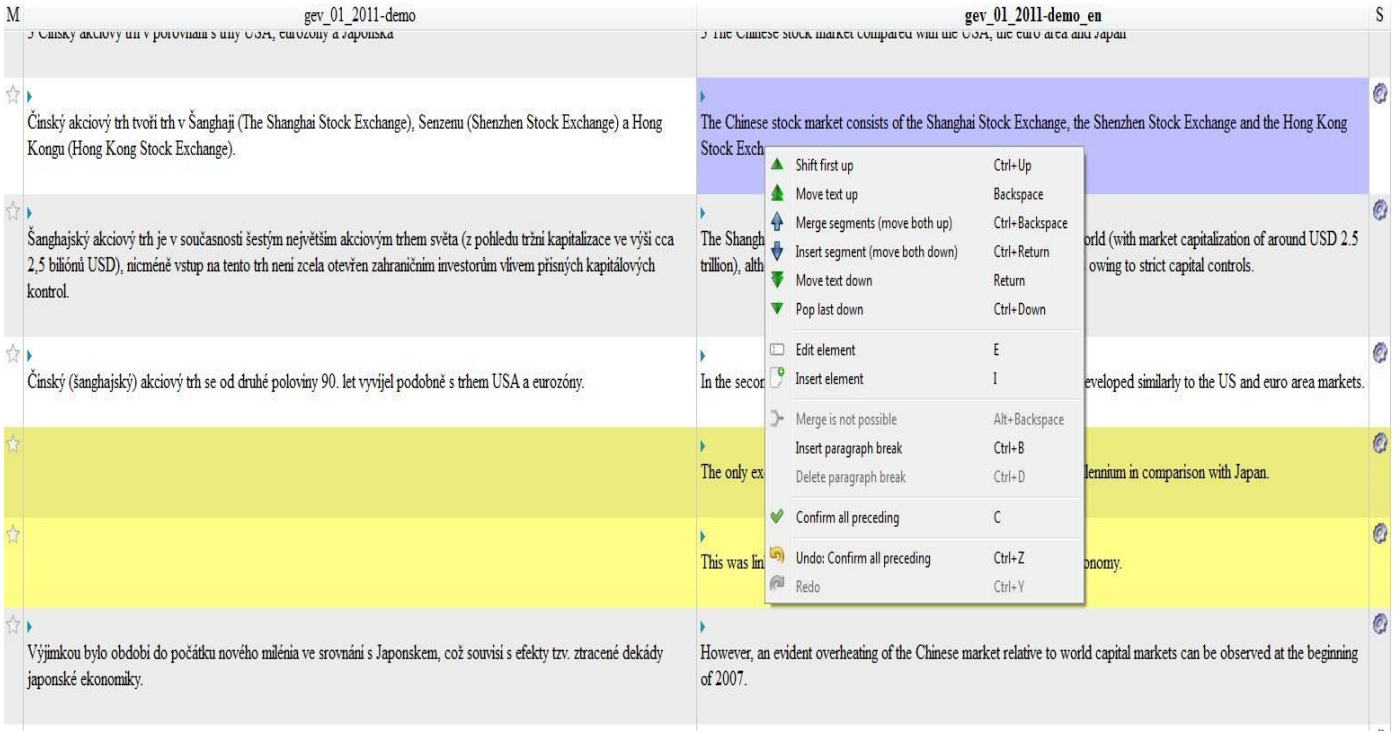
	M	gev_01_2011-demo	gev_01_2011-demo_en	S
1	☆	GLOBÁLNÍ EKONOMICKÝ VÝHLED – LEDENSekce měnová a statistiky Odbor vnějších ekonomických vztahů	GLOBAL ECONOMIC OUTLOOK – JANUARY Monetary and Statistics Department External Economic Relations Division	✓
2	☆	2011	2011	✓
3	☆	OBSAH 2	CONTENTS 2	✓
4	☆	OBSAH	CONTENTS 2	✓
5	☆	I SHRNUTÍ 3	I SUMMARY 3	✓
6	☆	II PŘEDPOVĚDI MEZINÁRODNÍCH INSTITUCÍ 4	II FORECASTS OF INTERNATIONAL INSTITUTIONS 4	✓
7	☆	II.1 HDP 4	II.1 GDP 4	✓
	☆	II.2 Porovnání předpovědi HDP a změna oproti předchozí předpovědi 5	<s> II.2 GDP forecast comparison and change from the previous forecast 5 </s>	

Obrázek č. 12:



	M	gev_06_2018	gev_06_2018_en	S
128			retaliatory tariffs on imports of US goods (such as soy and beef), as it hinted in March.	
129	☆	Počet nově vytvořených pracovních míst v nezemědělském sektoru v dubnu dosáhl 223 tisíc a míra nezaměstnanosti poklesla na 3,8 %, což je nejnižší úroveň od dubna roku 2000.	The US economy continues to expand.	
130	☆	Průměrná hodinová mzda vzrostla meziročně o 2,7 %.	The second estimate of GDP growth for 2018 Q1 was revised	
131	☆		downwards slightly to 2.2% (in quarter-on-quarter annualised terms), but growth in Q2 might exceed 4.5% according to the Atlanta Fed.	
132	☆		An improvement is expected in industry in particular; the ISM PMI leading	
133	☆		indicator suggests greater optimism of firms in all the main components (new orders, production and	
134	☆	Celková spotřebitelská inflace v květnu dále vzrostla na 2,7 % a jádrová PCE zůstává poblíž cíle centrální banky.	employment). Consumer sentiment remains robust, and retail sales went up by 5.9% year on year in May.	

Obrázek č. 13:



Obrázek č. 11 ukazuje, jak má ideálně vypadat zarovnání bez zásahů uživatele. Paralelní segmenty textu jsou zarovnány tak, že spolu sousedí český originál a překladatelský ekvivalent. Zcela na začátku každého zarovnání se nachází individuální název pro každou verzi zarovnání – tyto názvy mohou být libovolné a nemusí se shodovat s názvem celého dokumentu. U prvního segmentu vidíme, že je v obou verzích přeškrtnut. To je způsobeno tím, že text se nachází na stejném řádku, jako tag segmentu <s>. Tuto chybu ve formátování lze odstranit ručně. Zcela vlevo vidíme ikonu hvězdičky, která umožňuje zvýraznit segmenty červeně, můžeme si tak označit, že segment je pro nás důležitý apod. Úplně vpravo bývají segmenty označeny buď ozubeným kolečkem (segment je třeba ještě upravit) nebo odškrtnutím (potvrzeno, že segment není třeba upravit). Tyto funkce však nebyly pro zarovnáváníi v rámci této bakalářské práce vůbec spolehlivé. Při úpravě jednoho segmentu se totiž automaticky potvrdí všechny předcházející segmenty, i ty nesprávně zarovnané. Segmenty jsou editovatelné po dvojkliku myší. Pokud chceme vytvořit kvalitní paralelní korpus, manuální editace textu je nezbytná (Zanettin, 2013, 1-14.).

Obrázek č. 12 ukazuje častou situaci, kdy spolu dvě verze segmentů nekorespondují. Může se stát, že jedna věta se rozdělí na více segmentů nebo na jednotlivé řádky uvnitř segmentů (v programu pod názvem „elements“), v navazujícím textu je mezera, takže jsou některé segmenty zcela prázdné. Při vytváření předkládaného paralelního korpusu se u žádného

zarovnání nestalo, že by se verze segmentů stoprocentně shodovaly, vždy bylo potřeba provést dodatečnou editaci, ať už přepisováním, mazáním a kopírováním textu, nebo pomocí konkrétních funkcí programu InterText.

Na obrázku č. 13 jsou zobrazeny speciální funkce programu. Pokud chce uživatel zarovnání editovat, má na výběr z těchto možností:

a) Shift first up = první element v segmentu se posune o jeden segment výš
b) Move text up = celý segment se spojí se segmentem nad sebou
c) Merge segments = sousední segmenty (v obou verzích!) se posunou nahoru
d) Insert segments = sousední segmenty (v obou verzích!) se posunou dolů a zůstane po nich mezera
e) Move text down = celý segment se spojí se segmentem pod sebou
f) Pop last down = poslední element v segmentu se posune o jeden segment níž
g) Edit element = umožňuje upravit element stejně jako po dvojkliku
h) Insert element = přidává nový element do již existujícího segmentu
i) Merge both elements = spojí dva elementy dohromady
j) Insert paragraph break = vloží zalomení odstavce
k) Delete paragraph break = odstraní zalomení odstavce
l) Confirm all preceeding = označený segment se potvrdí zatržítkem tak jako všechny předcházející segmenty
m) Undo = Zpět
n) Redo = Vpřed

Porovnáme-li editace ***GEV ČNB*** z roku 2011 a z roku 2018, editace novějších verzí byly časově náročnější, bylo potřeba provést daleko více úprav (původní texty z tohoto roku obsahovaly více grafů, tabulek a měly jinak rozložené stránky se seznamem zkratek).

Pro vyexportování zarovnání jako dokument ve formátu ***XML*** je nutné v menu zvolit možnost „Export“. Při exportu tohoto typu dokumentu je třeba vybrat složku, kam budou umístěny tři soubory, a to jeden ***XML*** dokument pro každou jazykovou verzi, respektive jeden pro českou a jeden pro anglickou a třetím souborem bude ***XML*** dokument obsahující data o zarovnání těchto dvou verzí.

Program InterText nabízí možnosti exportu těchto typů souborů:

a) Newline aligned (one file)
b) Newline aligned (separate files)

c) ParaConc text (UTF-8)

d) ***TMX*** (stripped markup)

a) Obě verze textu jsou vyexportovány do jednoho TXT souboru, přičemž zarovnání je zaznamenáno formou tabulky oddělované tabulátorem a každý řádek textu vyjadřuje jednotlivý segment.

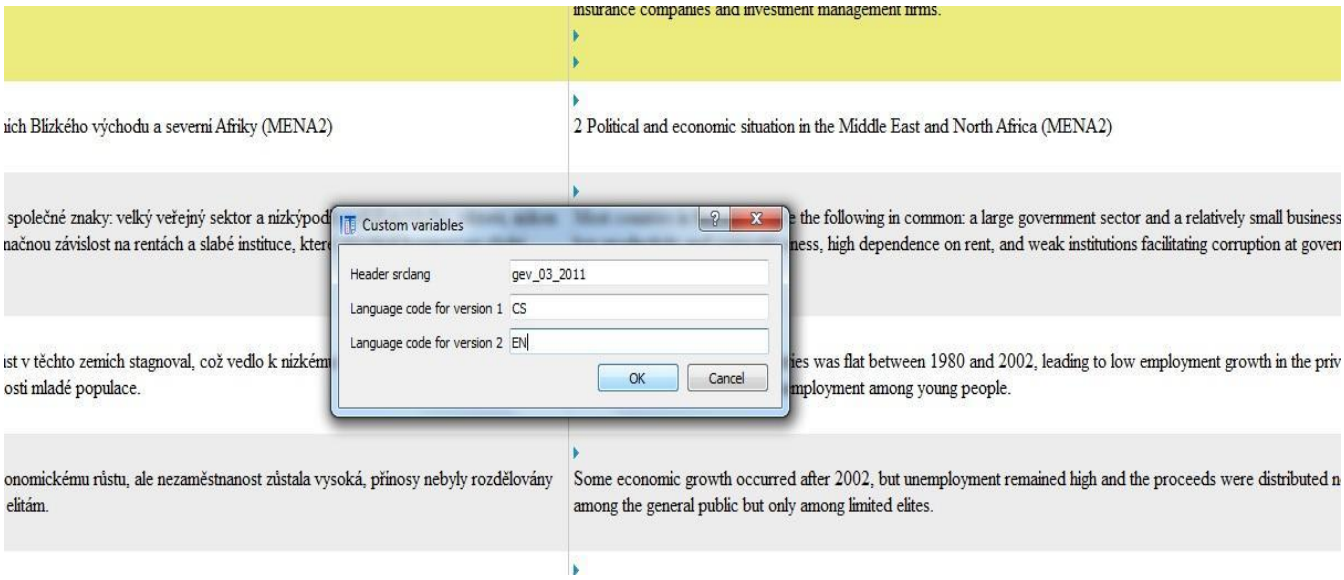
b) Každá z verzí textu je vyexportována v samostatném TXT souboru. Korespondující části textu z obou verzí se k sobě přiřadí podle příslušného čísla segmentu, které oba dva soubory obsahují.

c) Každá z verzí textu vyexportována jako samostatný ***XML*** soubor, ale od exportu tří ***XML*** souborů zmíněném výše se tato varianta liší v tom, že obě verze obsahují tzv. „seg“ tag. Tyto „seg“ tagy obsahují číselný údaj o tom, jak je text rozdělen v segmentech, a tak je možné při importu těchto souborů zajistit, že segmenty spolu budou korespondovat.

d) Vytvořená překladová paměť může být standardně využívána v CAT nástrojích. Při exportu ***TMX*** je nutné ke každé verzi přiřadit příslušný jazykový kód, tzn. „CS“ pro češtinu a „EN“ pro angličtinu. (viz následující obrázek) (Vondřička, 2016, 1-63.).[2]

Obr. č. 14: ***TMX*** – custom variables:

[2] Odkaz: https://wanthalf.saga.cz/repository/intertext/InterText_Editor-User_Guide.pdf. Datum stažení 25.5.2019. Pozn.: Na obrázcích týkajících se zarovnávání je zobrazen program InterText. Funkce programu jsou popsány v návodu, který se nachází na odkazu výše.



5. Možné využití korpusu

Pátá kapitola popisuje možnosti využití korpusu, vytvořeného v rámci předkládané bakalářské práce. Je zde zmíněn přínos v oblasti odborné terminologie, výzkumu mezijazykových systémových rozdílů a stylistické klasifikace.

Vytvořený paralelní korpus může být využit v oblasti překládání. Z hlediska odborné terminologie se jako jeho výhoda jeví obsažený seznam zkratek, který může překladateli objasnit, jaké zkratky jsou identické v cílovém i zdrojovém jazyce a jaké mají v každém jazyce jiný ekvivalent. Dále vidím přínos pro překlad kolokací a frázových sloves. Při překladu ekonomických výhledů a ekonomických textů obecně se překladatel musí vyhnout doslovnému znění překladu a více se soustředit na smysl. Cílem překladu je, aby čtenář pochopil smysl, nikoliv jednotlivá slova. Překladatel si musí dát pozor, aby nekopíroval syntaktické struktury zdrojového textu, protože text může být nesrozumitelný a stylisticky přeložený špatně (Munday, 2016, 29-47.).

Další možností, jak tento korpus využít, je výzkum v oblasti korpusové lingvistiky. Sestavený paralelní korpus by mohl posloužit výzkumu překladu českých sloves – na jeho základě lze např. udělat statistiku o tom, jak se v překladu ***GEV ČNB*** překládají slovesa: „uskutečnit“, „přebírat“, „kompenzovat“ a další a rovněž je možné zjistit, jak často a jaká se při překladu používají anglická frázová slovesa. Další možností je zkoumat, jak jsou přeloženy věty a souvětí obsahující spojky (a, neboť, nebo, ani, a proto atd.). Tento paralelní korpus umožňuje zkoumat systémové rozdíly mezi češtinou a angličtinou, konkrétně v žánru výhledových zpráv.

Potenciální výzkum je možné specifikovat na výhledové zprávy konkrétních institucí (***ČNB***), podle zdrojových jazyků textu a cílových jazyků při překladu. Porovnáme-li si výhledové zprávy z různých časových období, můžeme pozorovat, jak se měnil užívaný jazyk v čase.

Z lingvistického hlediska se také nabízí možnost zkoumat vlastnosti výhledových zpráv v oblasti funkčních stylů a stylistiky obecně. Lze zkoumat, do jakého funkčního stylu můžeme ***GEV ČNB*** zařadit a jak se prvky jednotlivých funkčních stylů prolínají. Výzkum systémových rozdílů mezi češtinou a angličtinou nemusí sloužit nutně pouze translatologům, ale má své využití i ve výzkumu lingvistiky obecně, například lze pozorovat, jaká morfologická nebo syntaktická pravidla se uplatňují.

Vzhledem k tomu, že se jedná o texty týkající se ekonomických prognóz, bylo by velmi přínosné zkoumat používání, časů a vidu u sloves. Systémovými rozdíly mezi oběma jazyky se zabývá Dagmar Knittlová v publikaci K teorii i praxi překladu vydané v roce 2000. Malý paralelní korpus ***GEV ČNB*** sestavený v rámci této bakalářské práce lze využít v případě výzkumů, které se zabývají tendencemi používání frázových sloves v češtině a angličtině. Dalšími oblastmi výzkumu, kde lze tento korpus využít, jsou: výskyt kolokací v češtině a angličtině, překlad kolokací, syntaktické rozdíly mezi češtinou a angličtinou, překlad textů obsahujících spojovací výrazy (angl. „linking words“) z angličtiny do češtiny, verbálnost a nominálnost v češtině a angličtině, specifika slovosledu v češtině a angličtině, funkční větná perspektiva atd. (Knittlová, 2000, 33-81.).

6. Závěr

Cílem této bakalářské práce bylo popsat, jak může vypadat průběh sestavování malého paralelního korpusu.

Dne 10. dubna 2019 jsem jako student Filozofické fakulty Univerzity Palackého oslovil ***ČNB*** s dotazem, jak je zajištěn překlad jí publikovaných globálních ekonomických výhledů z češtiny do angličtiny, konkrétně jestli si překládají výhledy autoři sami, nebo jestli jsou využívány služby externistů.

V reakci na dotaz zaslaný ***ČNB*** dorazila dne 15. dubna 2019 odpověď v tom smyslu, že anglickou verzi ***GEV ČNB*** nepíší samotní autoři – ***GEV ČNB*** překládají interní překladatelé s následnou korekturou externího korektora – rodilého mluvčího, ale autoři originálu na závěr kontrolují, jestli je překlad věcně správně. Také mi byl předán emailový kontakt na pana Jiřího Guta, vedoucího referátu publikačního a překladatelského. Z odpovědi ***ČNB*** vyplývá, že za účelem překladu využívají specializované pracoviště.

Koncepce ***GEV ČNB*** se v průběhu času změnila. Ve srovnání s verzemi z roku 2011 je ve verzích z roku 2018 více stran, méně textu, informace jsou zaznamenávány v daleko větší míře graficky, používají se jiné zkratky a ***GEV ČNB*** jsou doplněny o seznam příloh. Výhledy ***ČNB*** zvolené pro tuto práci se v porovnání s jinými výhledovými zprávami osvědčily jako dobrý zdroj informací.

GEV ČNB se ukázaly jako správně vybraný materiál, protože jsou kvalitně přeloženy a jejich strukturování je v rámci každého z vybraných roků stabilní (tzn. příliš se nemění struktura kapitol, formátování textu, seznam zkratek atd.). Důvodem je, že ***GEV ČNB*** mají v průběhu let konzistentní vzhled stran a strukturu informací (tj. obsah kapitol, formátování textu, seznam zkratek atd.). To nelze říct o ekonomických výhledech, které vydává Deloitte, ***MMF*** a ***Fed***. Byť mají výhledové zprávy společnosti Deloitte i českou verzi, pro sestavení korpusu jsou vhodnější materiály od ***ČNB***.

Při práci byl použit volně dostupný software na zarovnávání zvaný InterText. Aby bylo při práci s tímto softwarem dosaženo kýžených výsledků, je třeba provést precizní nastavení a zejména vzít v úvahu kódování souboru, pokud ho ukládáme ve formátu TXT – místo kódování ANSI je nutné nastavit kódování UTF-8, je nutné nastavit oddělování segmentů podle zalomení řádku a jaké elementy bude InterText vytvářet (viz Obr. č. 4). Je důležité zmínit, že pokud chceme paralelně zarovnat dva výchozí soubory a následně je exportovat do formátu ***TMX***, je třeba oddělit od sebe segmenty textu klávesou Enter již ve výchozích TXT souborech.

Během sestavování korpusu se ukázalo, že velkou výhodou je sestavovat korpus z elektronicky dostupných textů. Při extrakci textu z formátu PDF do jiných formátů se neosvědčilo používat online ***OCR*** nástroje, spolehlivější a méně časově náročné bylo vykopírovat relevantní úseky textu ručně. Neosvědčilo se zahrnutí úseků textu souvisejících s obsahem grafů, protože byly nutné dodatečné úpravy ve formátování a zarovnávání a tyto úseky textu po extrakci ztratily význam a relevanci.

Vytvořený korpus lze využít v oblasti ekonomického překladu, v oblasti slovní zásoby a kolokací, ale celý proces se ukázal~~a~~ jako časově velmi náročný, oproti původnímu očekávání zabral~~a~~ výrazně více času. Náročný~~á~~ byl~~a~~ zejména kvůli ručním opravám a úpravám textu, které bylo nutné provádět i v TXT souborech i v zarovnáváních nahraných v InterTextu. Jedno z možných řešení je používat v budoucnu ***OCR*** program, u kterého je rekognice znaků preciznější než u volně dostupných programů. Rovněž se zdá být vhodnější pracovat se zpoplatněným zarovnávacím programem.

7. Přílohy

Součástí příloh je i CD nosič, který obsahuje: české a anglické verze ***GEV ČNB*** z let 2011 a 2018 ve formátu PDF, příslušné TXT soubory a příslušné překladové paměti ve formátu ***TMX***, vytvořené exportem z programu InterText.

8. Summary

In my bachelor's thesis, I present my findings which are related to the practice of compilation of small parallel corpora and to the theoretical framework as well. The findings of my thesis indicate that the both language versions of global economic outlooks of the CNB comply with the theory that English is more analytic language than Czech. This is related to the issue of translation of phrasal verbs in global economic outlooks of ***CNB***. The systematic differences between Czech and English are described by Dagmar Knittlová in the book called "K teorii i praxi překladu" which was published in 2000. On 10th April 2019, I contacted the ***CNB*** via e-mail with a query. My query was how the translation of the economic outlooks is assured. In order to align the corpora, I used the InterText software.

Nevertheless, it is necessary to set the program precisely and also assume that the user has to set the UTF-8 encoding instead of ANSI in case of saving the TXT files. Next, it is necessary to set the "line break" as a form of separation. I received a response from the Czech National Bank on my query on 15th April 2019. The response said that the English versions of the particular economic outlooks are not written by the authors themselves, but the Czech versions are translated by the in-house translators. Subsequently, proofreading is done by a proofreader who is an externist and a native speaker of English. At the end, the authors of the original Czech versions check if the information provided in the translation is correct. On the basis of my interest in this topic, the ***CNB*** provided to me contact information to Mr Jiří Gut, who is the head of the publishing and translation department. Overall, the response from the ***CNB*** implies that this institution has a specialised department for translation.

The concept of global economic outlooks of the ***CNB*** has changed during the last several years. Comparing to the 2011 versions, the 2018 versions contain less text, more charts and figures, different abbreviations and annexes are added. It has been proven that the economic outlooks of the ***CNB*** are a good source of information. The corpus compiled by me can be possibly used within the area of economic translations, it is beneficial when researching vocabulary and collocations. I find the global economic outlooks of the ***CNB*** a very good choice for my bachelor's thesis. The reason is that unlike the outlooks of the Deloitte company, IMF and ***Fed***, the ***CNB*** outlooks have consistent layout and structure during all the years (i.e. the content of chapters, text formatting, lists of abbreviations etc.).

When I was compiling my corpus, I found out that it is a big advantage to compile a corpus from texts which are electronically available. When I was extracting the text from the PDF format to a different one, I found out that it is better to extract the relevant segments of text

manually than to do it with the help of ***OCR*** tools which are available online. I also found out that it is not a good idea to extract the segments of text related to figures and charts because it required very time-consuming adjustments of formatting and alignment. In addition, such segments of text lost their meaning and relevance after being extracted. To sum up the theses, it was very difficult to compile the corpus, mainly because of manual corrections of the text extracted which were necessary in the TXT files and also within the InterText alignment interface. One of possible solutions I suggest to make the workflow more effective is to choose a better ***OCR*** program (probably a program which is not available online for free). Such ***OCR*** program should have much better output of recognition of characters. In addition, it would be a good strategy to purchase a computer program which is not a freeware so that the quality of alignment may be better. My thesis cannot be used only for the scientific purposes, not for the commercial purposes. The use and creation of this bachelor's thesis is regulated by the provisions of Law No. 89/2012 Coll. concerning scientific licences.

9. Použité zdroje

Deloitte. „Czech Economic Outtlook for 2018: Controlled Slowdown," *Czech Economic Otlook for 2018,* prosinec 2017.

Deloitte. „2019 Banking and Capital Markets Outlook: Reimaging transformation," *Deloitte Centre For Financial Services,* září 2018.

Evropská unie. „European Economic Forecast: Europe 2019 (Interim)," „*European Economic Forecast,* únor 2019.

Fowler, W. Chauncey. *English grammar: The English language in its elements and forms; with a history of its origin and development; designed for use in colleges and schools.* New York, USA: Harper & Brothers, Publishers, 1850.

Frankenberg-Garcia, Ana a Santos, Diana. „Introducing COMPARA, the Portuguese-English parallel translation corpus." In: *Corpora in Translation Education,* editovali Federico Zanettin; Silvia Bernardini; Dominic Stewart. 71-72. Manchester: St. Jerome Publishing, 2003.

Grishman, Ralph. *Computational Linguistics: An Introduction (Studies in Natural Language Processing).* Cambridge, VB: Cambridge University Press, 1986.

Gurría, Angel. Growth has peaked amongst escalating risks. Ekonomický výhled, OECD, Paříž, 21.listopadu 2018.

Holubová, Irena a Pokorný, Jaroslav a Richta, Karel a kol. *XML technologie: principy a aplikace v praxi*. Praha: Grada, 2008.

Chlumská, Lucie. „Není korpus jako korpus: Korpusy v kontrastivní lingvistice a translatologii." In: *Časopis Pro Moderní Filologii*, č. 2 (2014): 221-32.

International Monetary Fund, „World Economic Outlook: Growth Slowdown, Precarious Recovery," *World Economic Outlook: Growth Slowdown, Precarious Recovery,* duben 2019.

Kennedy, Graeme. *An introduction to corpus linguistics*. Abingdon-on-Thames, VB, New York, USA: Routledge, 2014.

Knittlová, Dagmar. *K teorii i praxi překladu*. Olomouc, Česká republika: Univerzita Palackého, 2000.

Kováříková, Dominika. „Kvantitativní charakteristiky termínů." Dizertační práce, Univerzita Karlova v Praze, Filozofická Fakulta, 2014.

Mareš, Petr. *Úvod do lingvistiky a lingvistické bohemistiky*., Praha, Česká republika: Univerzita Karlova v Praze, 2014.

Marco, Josep a van Lawick, Heike. „Using corpora and retrieval software as a source of materials for the translation classroom.“ In: *Corpus Use and Translating. Corpus use for learning to translate and learning corpus use to translate,* editováno Allison Beeby, Patricia Rodríguez Inés, Pilar Sánchez-Gijón. 10. Castellón de la Plana: Universitat Jaume I, 2009.

Martinková, Michaela. „K metodologii využití paralelních korpusů v kontrastivní lingvistice.“ In: *Naše řeč,* č. 4-5 (2014): 270-285.

Minářová, Eva. *Stylistika pro žurnalisty.* Praha, Česká republika: Grada Publishing, 2011.

Munday, Jeremy. *Introducing Translation Studies: Theories and Applications.* Abingdon-on-Thames, New York: Routledge, 2016.

Olohan, Maeve. „Leave It Out! Using a Comparable Corpus to Investigate Aspects of Explicitation In Translation.“ In: *Cadernos de Tradução,* č. 9 (2002): 153-169.

Scott, Juliette. „Can Genre-Specific DIY corpora, Compiled by Legal Translators Themselves, Assist them in 'Learning the Lingo' of Legal Subgenres.“ In: *Comparative Leglinguistics,* č. 12/2012 (2012): 87-99.

Pearson, Jennifer. Using parallel texts in the translator training envieronment. In: *Corpora in translator education,* editovali Federico Zanettin; Silvia Bernardini; Dominic Stewart. 15-24. Manchester a Northampton: St. Jerome Publishing, 2003.

Pořízka, Petr. *Tvorba Korpusů a vytěžování jazykových Dat: Metody, Modely, nástroje*., Olomouc, Česká republika: Univerzita Palackého v Olomouci,– 2019.

Ptatscheková, Jitka a Dittrichová, Jaroslava. *Dvacet let české koruny na pozadí vývoje obchodního bankovnictví v České republice.* Praha: Grada Publishing, a.s., 2013.

Saldanha, Gabriela a O'Brien, Sharon. *Research Methodologies in Translation Studies.* Abingdon-on-Thames, New York: Routledge, 2014.

Marco, Josep a van Lawick, Heike. „Using corpora and retrieval software as a source of materials for the translation classroom.“ In *Corpus Use and Translating. Corpus use for learning to translate and learning corpus use to translate*, editováno Allison Beeby, Patricia Rodríguez Inés, Pilar Sánchez-Gijón. 10. Castellón de la Plana: Universitat Jaume I, 2009.

Frankenberg-Garcia, Ana a Santos, Diana. „Introducing COMPARA, the Portuguese-English parallel translation corpus.“ In *Corpora in Translation Education*, editovali Federico Zanettin; Silvia Bernardini; Dominic Stewart. 71-72. Manchester: St. Jerome Publishing, 2003.

Pearson, Jennifer. „Using parallel texts in the translator training envieronment.“In *Corpora in translator education*, editovali Federico Zanettin; Silvia Bernardini; Dominic Stewart. 15-24. Manchester a Northampton: St. Jerome Publishing, 2003.

Gurría, Angel. „Growth has peaked amongst escalating risks.“ Ekonomický výhled, OECD, Paříž, 21.listopadu 2018.

World Bank Group. G*lobal Economic Prospects: Darkening Skies*, leden 2019.

Zanettin, Frederico. „Corpus Methods for Descriptive Translation Studies.“ *V International Conference on Corpus Linguistics*(2013): 1-14.

Deloitte. „Czech Economic Outtlook for 2018: Controlled Slowdown,“ *Czech Economic Outlook for 2018*, prosinec 2017.

Deloitte. „2019 Banking and Capital Markets Outlook: Reimaging transformation," *Deloitte Centre For Financial Services*, září 2018.

Zehnalová, Jitka, „Převod stylu v literárním překladu: využití paralelních korpusů a dalších elektronických nástrojů." In: *Acta Universitatis Carolinae Philologica*, 2. číslo (2018): 101–105.

Online zdroje:

Burnard, Lou. The British National Corpus (BNC). „What Is the BNC?" Citováno 15. února 2019. http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro.

Davies, Mark a kol. The Corpus of Contemporary American English (***COCA***). „Corpus of Contemporary American English." Citováno 16. února 2019. https://corpus.byu.edu/coca/.

Evans, David. „Corpus building and investigation for the Humanities: An on-line information pack about corpus investigation techniques for the Humanities", *Unit 2: Compiling a corpus*: 1-5. Citováno 1. května 2019. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/Intro/Unit2.pdf

Filozofická fakulta Masarykovy univerzity. „Parallel Corpus of English and Czech Texts. Rambousek Jan a Chamonikolasová, Jana a kol. „KACENKA." Citováno 21. února 2019. http://www.phil.muni.cz/angl/kacenka/kachna.html

Filozofická fakulta Univerzity Karlovy. Křen, Michal. „pojmy: korpus: Korpus a jeho využití: Využití v lingvistice." Citováno 12. února 2019. http://wiki.korpus.cz/doku.php/pojmy:korpus#vyuziti_v_lingvistice.

Filozofická fakulta Univerzity Karlovy. Křen, Michal. „pojmy: token.“ Citováno 13. února 2019. https://wiki.korpus.cz/doku.php/pojmy:token.

Filozofická fakulta Univerzity Karlovy. Křen, Michal. „pojmy: synchronní.“ Citováno 15. února 2019. https://wiki.korpus.cz/doku.php/pojmy:token.9.

Filozofická fakulta Univerzity Karlovy. Křen, Michal. „pojmy: diachronní.“ Citováno 15. února 2019. https://wiki.korpus.cz/doku.php/pojmy:diachronni.

The British National Corpus (BNC). Burnard, Lou. „Design of the Corpus.“ Citováno 12. února 2019. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

HarperCollins Publishers L.L.C. „The History of COBUILD.“ Citováno 15. února 2019. https://www.collinsdictionary.com/cobuild/.

IBM. „Pioneering the Computational Linguistics and the Largest Published Work of All Time.“ Citováno 12. února 2019. https://web.archive.org/web/20120327122219/http://www.ibm.com/ibm100/it/en/stories/linguistica_computazionale.html.

Rosen, Alexandr a kol. Filozofická fakulta Univerzity Karlovy v Praze. „Intercorp.“ Citováno 6. května 2019. ucnk.korpus.cz/intercorp/

Kilgariff, Adam a Rychlý, Pavel a kol.„OPUS parallel corpus | Sketch Engine“ citováno 6. května 2019. https://www.sketchengine.eu/opus-parallel-corpora-2/

Komárek, Luboš a kol. „Globální ekonomický výhled - Česká národní banka.“ Citováno 17. června 2019. https://www.cnb.cz/cs/menova-politika/gev/index.html

„FREE Online OCR - Convert PDF to Word or Image to text.“ Online OCR, zobrazeno 10. března 2019. https://www.onlineocr.net/

Vondřička, Pavel. *InterText editorv1.5 comprehensive guide.* Univerzita Karlova, Filozofická fakulta. Ústav Českého národního korpusu, Česká republika, Praha, 2002.

10. Abstract

In my thesis, I compile a small parallel corpus of the global economic outlooks issued by the Czech National Bank during the years 2011 and 2018. The aim of the thesis is to describe and experience the process of creating small parallel corpora and explore its potential use.Within the content of this thesis, there is put emphasis on both linguistic and extralinguistic aspects of the process of creating a small parallel corpus. The extralinguistic information are included mainly in the fourth chapter which characterises outlooks as a text type.

However, the content provided in the thesis is both theoretical and it also provides the reader with description of some practical aspects of the process of creating of a parallel corpus. In terms of the theoretical information, this thesis describes the historical background of evolution of corpus linguistics, possible ways of classification of corpora and there are mentioned important linguistic scientists and specific examples of corpora. Regarding the practical aspects of the process of creating of parallel corpora, I describe how I extracted the content from the original texts (global economic outlooks mentioned above), how I worked with the InterText software, characteristics of the alignment process and my system— of giving names to the alignment computer files. The fifth chapter of my thesis describes the possibilities of use of my corpus.

The conclusion of this thesis describes my findings during the creation process of my corpus and it is emphasised there that the process of a small parallel corpus is very time-consuming, considering the extraction process, OCR, alignment and additional editing and corrections of the content extracted.

Key words:

corpus linguistics, parallel corpus, alignment, outlooks, global economic outlook, Czech National Bank, Czech National Corpus, lemmatisation

11. Anotace

V rámci předkládané bakalářské práce byl sestaven malý paralelní korpus globálních ekonomických výhledů České národní banky, které byly vydány v letech 2011 a 2018. Tato bakalářská práce popisuje jak lingvistické, tak i extralingvistické aspekty sestavování malého paralelního korpusu. Extralingvistické informace jsou v této práci představeny zejména ve čtvrté kapitole, kde jsou výhledové zprávy charakterizovány jako typ textu.

Předkládaná práce obsahuje informace teoretického rázu a zároveň osvětluje některé z praktických aspektů sestavování paralelního korpusu. Co se týče teoretického obsahu, popisuje tato práce historický vývoj korpusové lingvistiky, možnosti klasifikace korpusů a také jsou zde zmíněni významní lingvisté a konkrétní korpusy. Mezi popisované praktické aspekty procesu sestavování malého paralelního korpusu patří extrakce obsahu z původních textů (***GEV ČNB***), práce s programem InterText, charakteristika procesu zarovnávání a systém pojmenovávání dokumentů, které souvisí se zarovnáváním.

Možným využitím sestaveného korpusu se zabývá pátá kapitola této práce.

Závěr této práce shrnuje zjištění získaná během procesu sestavování. V závěru je zdůrazněna velká časová náročnost sestavování malého paralelního korpusu. Ta zvláště vynikne, uvážíme-li, jak probíhala extrakce obsahu z originálních textů, dále i to, jaký mělo průběh ***OCR***, zarovnávání a také dodatečné editace a opravy extrahovaného obsahu.

Klíčová slova:

korpusová lingvistika, paralelní korpus, zarovnávání, výhledové zprávy, globální ekonomický výhled, ČNB, Český národní korpus, lemmatizace