2. FRAZEOLOGIE NIZOZEMŠTINY A PRÁCE S KORPUSY

Mgr. Kateřina Křížová, Ph. D. (katerina.krizova@upol.cz, kamalkova@hotmail.com)

2.1. Co je jazykový korpus?

Jazykový korpus (*corpus*) je *„rozsáhlý soubor autentických textů (psaných nebo mluvených) převedený do elektronické podoby v jednotném formátu tak, aby v něm bylo možné jednoduše vyhledávat jazykové jevy, zejména slova a slovní spojení neboli kolokace. Korpus zobrazuje jazykové jevy v jejich přirozeném kontextu, a umožňuje tak vytvářet na základě reálných dat podložený jazykový výzkum v rozsahu, který byl dříve nemyslitelný.“*[1]

Korpusový výzkum je v současné době převládající metodologií v lexikologii a lexikografii, uplatňuje se však i v jiných rovinách popisu jazyka. Většina korpusů nizozemského jazyka je dostupná pro výzkumné účely zdarma, a to prostřednictvím *TST-Centrale (*http://tst-centrale.org/nl/). Od roku 2016 však přecházejí postupně digitální materiály nabízené přes TST-Centrale pod správu **Institutu pro nizozemský jazyk** (*Instituut voor de Nederlandse Taal – INT*).[2]

2.2 Základní terminologie

Jazykové korpusy obsahují většinou texty různě označkované neboli **anotované** (*geannoteerd*), tj. opatřené metalingvistickými informacemi. Jedná se jednak o informace o samotných textech (autor, původ, doba vzniku apod.), jednak o doplňující lingvistické informace k jednotlivým textovým slovům a jazykovým jevům.

Nejmenší jednotkou v textu korpusu je tzv. **token** (*token*), což je většinou grafické slovo (tj. řetězec znaků oddělený mezerou v textu), resp. jedna jeho konkrétní realizace, může jím však být např. i interpunční znaménko.[3] Z tohoto důvodu se velikost korpusu udává většinou v tokenech, příp. v textových či grafických slovech. Členění textu na tokeny se nazývá **tokenizace** (*tokenizing*). Při tomto procesu jsou od sebe oddělovány rovněž jednotky, které v původním textu sice tvoří jeden grafický celek, ale nepatří k sobě, např. interpunční znaménka psaná za slovy. **Korpusová lingvistika** (*corpustaalkunde, corpuslinguïstiek*) pracuje rovněž s pojmem **typ** (*type*), což je jednotka abstrakce. Na webových stránkách *Českého národního korpusu* (ČNK) je tento termín specifikován následovně: *„Zatímco tokenem se míní vždy konkrétní realizace jednotky (konkrétní výskyt formy) v určitém kontextu, typ je jednotka dekontextualizovaná (na kontextu nezávislá), která je schopna nabývat*

---

[1] Viz definice pojmu *korpus* na webové stránce *Českého národního korpusu*: https://wiki.korpus.cz/doku.php?id=pojmy:korpus&rev=1416829573.

[2] Institut pro nizozemský jazyk (*Instituut voor de Nederlandse Taal*) má zaujmout výsadní postavení v péči o jazyk v nizozemské jazykové oblasti. Aktuální informace viz webové stránky INT: http://ivdnt.org/, aktuální seznam lexikonů, korpusů a dalších digitálních materiálů nabízených prostřednicvím INT viz tamtéž.

[3] Srov. Boonen & Harmes (2013: 225) a definice pojmu *token* na webových stránkách *Českého národního korpusu*: http://wiki.korpus.cz/doku.php/pojmy:token. O jednotlivých tokenech se někdy mluví také jako o tzv. *pozicích*.

takových vlastností, jako je frekvence (...)".[4] V této souvislosti se nejčastěji se uvažuje o **slovním tvaru**, který je v korpusové lingvistice označován anglickým termínem **word,** přičemž většinou rozhoduje i velikost písmen – např. *schrijf* a *Schrijf* jsou v korpusech považovány za dva slovní tvary, dva typy, přičemž typ *schrijf* může mít mnoho různých realizací, tj. tokenů. Na druhou stranu např. u nizozemského slovního tvaru *je* není rozlišováno, zda se jedná o tvar osobního zájmena (ty) či o tvar přivlastňovacího zájmena (tvůj).

Při tzv. **lemmatizaci** (*lemmatisering*) je každému tokenu (tj. konkrétní realizaci grafického slova) v korpusu přiřazeno **lemma** (*lemma*), tj. základní slovníková podoba hesla (u nizozemských substantiv tvar v singuláru, u adjektiv tvar bez koncovky *e*, u sloves tvar infinitivu). Lemmatizace je součástí automatické **morfologické anotace** (*morfologische annotatie*). Morfologické značky, které se používají v korpusové lingvistice a zahrnují souhrn gramatických, příp. dalších informací, se nazývají **tagy** (*tags*). **Tagování** (*tagging*) je přiřazení tagu jednotlivým tokenům, tj. pozicím. **Parts of speech tagging** znamená přiřazení značky slovního druhu jednotlivým tokenům. Pro korpus, který obsahuje **syntaktickou anotaci** (*syntactische annotatie*), se užívá označení *treebank* (tzv. stromová databanka), neboť se v něm naznačuje struktura věty v podobě závislostního stromu[5] (viz obrázek č. 1).

Obrázek č. 1: Ukázka ze syntakticky anotovaného korpusu *Lassy Groot-corpus*[6]

Důležitým empirickým instrumentem je při korpusové analýze tzv. **konkordance** (*concordantie*). Konkordance zahrnuje *„všechny doklady (výskyty) hledaného jevu v korpusu spolu s okolním kontextem. V praxi se v rámci konkordance rozlišuje **KWIC (tj. Key word in context)**, tedy hledané slovo/jev a jeho pravý a levý kontext. Jeden řádek konkordančního seznamu se označuje jako*

[4] Viz definice pojmu *typ* a *slovní tvar*: http://wiki.korpus.cz/doku.php/pojmy:typ a http://wiki.korpus.cz/doku.php/pojmy:word.

[5] Závislostní strom je zvláštní případ orientovaného grafu, který je výsledkem syntaktické analýzy věty. Proces, kdy je větám automaticky přiřazována syntaktická struktura, se v korpusové lingvistice nazývá parsing. Počítačové programy, které provádějí automatickou syntaktickou analýzu vět, se nazývají **parsery** – podrobnější informace k těmto pojmům viz: https://wiki.korpus.cz/doku.php/pojmy:parser.

[6] Viz: http://www.let.rug.nl/~vannoord/alp/Alpino/TreebankTools.html

konkordanční řádek.“[7] Pro ilustraci uvádíme na obrázku č. 2 příklad konkordančních řádků při vyhledávání slovního tvaru *schrijf* (= KWIC) v korpusu *Corpus Hedendaags Nederlands*.

Left context	Hit text	Right context	Lemma	Part of speech
Elke leeslijster zingt zoals hij gebekt is (2005)				
... dat ik alleen over problemen	**schrijf**	, maar daar ben ik nu ...	schrijven	VRB(finiteness=fin,mood=imp|ind,tense=pres,number=sg)
... zichzelf in mijn romans. Ik	**schrijf**	boeken die mijn lezers in ...	schrijven	VRB(finiteness=fin,mood=imp|ind,tense=pres,number=sg)
Blijf praten met je kind (2005)				
... dingen die wel goed lopen.	**Schrijf**	een brief of stuur een ...	schrijven	VRB(finiteness=fin,mood=imp|ind,tense=pres,number=sg)
De Standaard (2005)				
... schrijfproces een ontdekkingstocht: „Als ik	**schrijf**	, ben ik enorm geconcentreerd. Ik ...	schrijven	VRB(finiteness=fin,mood=imp|ind,tense=pres,number=sg)
Tom Boonen: „Ik gun het Robbie" (2005)				
... in Armstrongs laatste Tour. Zo	**schrijf**	ik mee een stukje Tour-geschiedenis ...	schrijven	VRB(finiteness=fin,mood=imp|ind,tense=pres,number=sg)

Obrázek č. 2: Slovní tvar *schrijf* v korpusu *Corpus Hedendaags Nederlands*

Pro vytěžování dat z korpusů a usnadnění práce s korpusy se používají speciální počítačové programy, tzv. **korpusové manažery** (*corpus managers*), které umožnují nejen různé typy vyhledávání v korpusu a prohlížení výsledků, ale i filtrování a jednoduché statistické vyhodnocování nalezených dat.

Při práci s jazykovými korpusy je však třeba mít na paměti, že kvalita anotací se v jednotlivých korpusech liší – pokud anotace probíhá pouze automaticky, je třeba počítat s větším procentem chyb než při manuální anotaci korpusu. Také je třeba vždy nejprve dobře zvážit, který z dostupných korpusů je pro zkoumání sledovaného jazykového jevu nejvhodnější – čím specifičtější, neobvyklejší jazykový jev zkoumáme, tím větší, příp. specifičtější korpus je třeba mít k dispozici. Přitom si je zároveň třeba uvědomit, že pokud vybraný korpus zkoumaný jazykový jev nezachycuje, neznamená to nutně, že daný jev neexistuje.

2.3 Typy jazykových korpusů

Jazykové korpusy je možné dělit např. na reprezentativní a nereprezentativní, na obecné (referenční) a specializové, na psané a mluvené, na synchronní a diachronní nebo z hlediska počtu jazyků na jednojazyčné a vícejazyčné.

Referenční korpusy (*referentiecorpora*) jsou velmi obsáhlé a měly by zaznamenávat běžný jazykový úzus, třebaže žádný korpus nemůže zachytit celou šíři používání přirozeného jazyka. Za reprezentativní jsou označovány, obsahují-li v dostatečném množství texty všech variet daného jazyka. Pokud v nich jsou zastoupeny texty různých jazykových stylů a poměr mezi jednotlivými jazykovými varietami v korpusu odpovídá poměru, v jakém se variety vyskytují v běžném úzu, označují se za **vyvážené** (*gebalanceerd*).[8] Pojmy *reprezentativnost* a *vyváženost* jazykového korpusu však v korpusové lingvistice doposud nejsou jednoznačně definovány. Příkladem pokusu o referenční korpus současného psaného nizozemského jazyka je *ANW-corpus*[9] s více než 100 miliony slov, z něhož

[7] Viz definice pojmu *konkordance*: https://wiki.korpus.cz/doku.php/pojmy:konkordance (srov. též Boonen & Harmes 2013: 228–229).

[8] Srov. Boonen & Harmes (2013: 224) a definice základních pojmů z korpusové lingvistiky dostupné on-line na webových stránkách *Českého národního korpusu*: https://wiki.korpus.cz/doku.php/pojmy:prehled_pojmu.

[9] Viz: http://anw.inl.nl/. Podrobněji k zastoupení textů v korpusu ANW viz: http://anw.inl.nl/anwcorpus a https://www.sketchengine.co.uk/algemeen-nederlands-woordenboek-anw-corpus/.

vychází elektronický slovník *Algemeen Nederlands Woordenboek*, či korpus *SoNaR*, který obsahuje přibližně 500 milionů slov a je založen na textech z období let 1954–2002.[10]

Specializované korpusy (*speciale corpora*) se zaměřují na specifické užití jazyka, na jistý typ textů, takže nepokrývají celou šíři používání jazyka. Příkladem mohou být různé korpusy novinových článků – pro nizozemštinu např. *27-Miljoen-Krantencorpus*. Literární texty z období let 1250–1500 zahrnuje *Corpus Middelnederlands*.

Většina korpusů patří mezi **korpusy psané** (*geschreven corpora*), neboť zpracování mluveného jazyka je mnohem obtížnější. Pro tvorbu psaných korpusů je k dispozici velké množství psaných textů, v dnešní době často již přímo dostupných v digitální podobě (noviny a časopisy, knihy, zákony, odborné publikace, elektronická komunikace apod.). Digitalizovány jsou postupně i starší psané dokumenty, a to včetně historického materiálu. Velmi rozsáhlý je psaný korpus *Corpus Hedendaags Nederlands*[11], který vznikl spojením korpusů *5 Miljoen Woorden Corpus*, *27 Miljoen Woorden Corpus*, *38 Miljoen Woorden Corpus* a *PAROLE Corpus* a doplněním o novinové články z nizozemských deníků NRC a De Standaard. Obsahuje více než 800 000 textů z období let 1814–2013, a to především publicistického stylu a odborného právnického stylu.

Mluvené korpusy (*gesproken corpora*) jsou založeny na nahrávkách a transkripcích mluveného jazyka, takže zpravidla nejsou tak rozsáhlé jako korpusy psané. Největším a nejznámějším korpusem mluveného jazyka pro nizozemštinu je *Corpus Gesproken Nederlands*, jenž obsahuje přibližně 900 hodin nahrávek (přibližně 9 milionů slov) z období let 1998–2003, a to mluvčích z Nizozemska i Vlámska. Zvukové nahrávky spontánních i připravených mluvených projevů jsou doplněny transkripcemi (mj. fonetickou a ortografickou) a texty jsou opatřeny anotací (slovnědruhovou a syntaktickou).[12]

Synchronní korpusy (*synchrone corpora*) zaznamenávají jazyk jednoho konkrétního (většinou úzce vymezeného) časového období. Ve slovníku základních pojmů z korpusové lingvistiky, který je k dispozici uživatelům ČNK, se k tomu upřesňuje: „*Z pohledu současného jazyka se jako synchronní jeví korpus, který zachycuje jazyk živý, tj. takový, který je užíván žijícími mluvčími*.“[13]

Diachronní korpusy (*diachronne corpora*) jsou soubory více korpusů z různých dob vývoje jazyka, které nabízí možnost zkoumání jazykového vývoje.

Vícejazyčné korpusy (*meertalige corpora*) zahrnují texty dvou (či více) různých jazyků a patří mezi ně např. paralelní korpusy a srovnatelné korpusy. **Paralelní korpusy** (*parallele corpora*) obsahují stejné texty v různých jazykových verzích. Jako příklad vícejazyčného paralelního korpusu lze pro nizozemštinu uvést *Dutch Parallel Corpus,* který je dostupný pro dvojice jazyků nizozemština – angličtina a nizozemština – francouzština, a to oboustranně, přičemž texty jsou anotované a

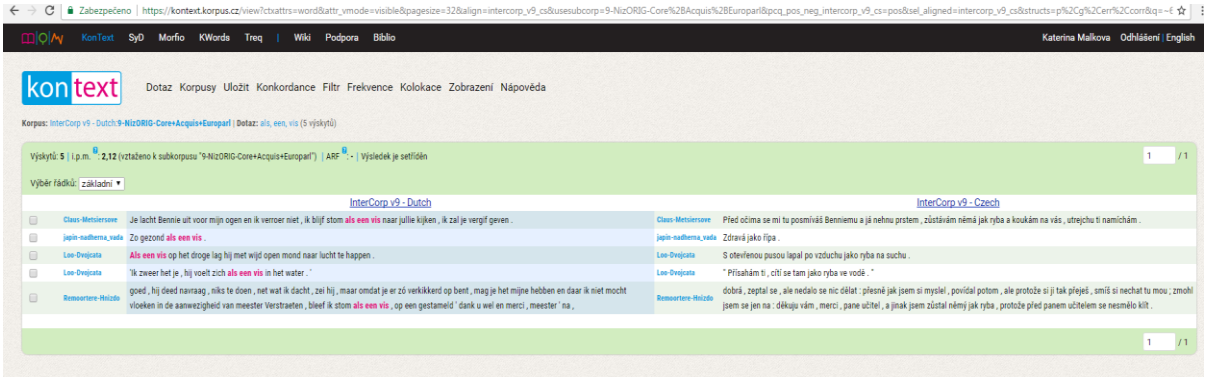
[10] Korpus *SoNaR* je dostupný od roku 2015 on-line přes portál INL CLARIN. Podrobnější charakteristika korpusu *SoNaR* viz: http://taalunieversum.org/inhoud/corpora/d-coi-en-sonar či http://tst-centrale.org/nl/tst-materialen/corpora/sonar-corpus-detail. Více o celém projektu viz: http://lands.let.ru.nl/projects/SoNaR/description.html.

[11] Dostupný na: http://chn.inl.nl/.

[12] Viz: http://tst-centrale.org/nl/tst-materialen/corpora/corpus-gesproken-nederlands-detail.

[13] https://wiki.korpus.cz/doku.php/pojmy:synchronni.

zarovnané (*gealigneerd*) na úrovni vět.[14] Dalším příkladem vícejazyčného paralelního korpusu je korpus *Europarl* vycházející ze zápisů jednání Evropského parlamentu. Nizozemština je zastoupena rovněž jako jeden z 39 jazyků ve vícejazyčném paralelním korpusu *InterCorp*[15], v němž má každý cizojazyčný text svou českou verzi (ať už se jedná o originál či o překlad) (viz obrázek č. 3). Korpus *InterCorp* je součástí ČNK a skládá se ze dvou částí – z tzv. jádra, které je tvořeno ručně zarovnanými převážně beletristickými texty, a z kolekce[16], což jsou texty získané ve více jazycích, které jsou automaticky zpracované a zarovnané. Nizozemské texty jsou opatřeny značkami, nejsou však lemmatizovány, což omezuje možnosti využití korpusu *InterCorp* pro korpusovou analýzu (Křížová 2015).[17]



Obrázek č. 3: Slovní spojení *als een vis* v paralelním korpusu *InterCorp*

Jiným typem vícejazyčných korpusů jsou **srovnatelné korpusy** (*vergelijkbare corpora*). Tyto korpusy na rozdíl od paralelních korpusů neobsahují originální texty a jejich překlady do různých jazyků, nýbrž *„texty vybrané podle týchž kritérií (žánru, zaměření, délky apod.) v různých jazycích (příp. v různých varietách jednoho jazyka)“* Chlumská (2014: 225). Nizozemské texty jsou součástí např. souboru nereferenčních srovnatelných webových korpusů *ARANEA*[18], které jsou zdarma dostupné na webových stránkách ČNK (viz: www.korpus.cz).

2.4 Nizozemská frazeologie v různých typech korpusů (praktické ukázky vyhledávání)

2.4.1 Paralelní korpus InterCorp

2.4.2 Korpus a slovník ANW

2.4.3 Corpus Hedendaags Nederlands

---

[14] Více informací a přístup k tomuto korpusu nabízí TST-centale na: http://tst-centrale.org/nl/tst-materialen/corpora/dutch-parallel-corpus-niet-commercieel-detail.

[15] V září roku 2016 byla zpřístupněna již 9. verze korpusu *InterCorp* (viz: https://wiki.korpus.cz/doku.php/cnk:intercorp a http://wiki.korpus.cz/doku.php/cnk:intercorp:verze9).

[16] Kolekci paralelního korpusu InterCorp tvoří jednak publistické články z webových stránek *Project Syndicate* a *Presseurop*, jednak zápisy jednání Evropského Parlamentu z let 2007–2011 z korpusu *Europarl*, právní texty Evropské Unie z korpusu *Acquis Communautaire* a filmové titulky z databáze *Open Subtitles* (více viz: https://ucnk.ff.cuni.cz/intercorp/?req=page:info).

[17] Korpus je po registraci a přihlášení dostupný pro nekomerční účely zdarma na: www.korpus.cz.

[18] Podrobnější informace ke korpusům ARANEA viz: http://ucts.uniba.sk/aranea_about/index.html.

Citováno z:

KŘÍŽOVÁ, Kateřina, KLUKOVÁ, Markéta, BOSSAERT, Benjamin, HORST, Pim van der & Wilken ENGELBRECHT (2017): *Capita selecta z nizozemské lingvistiky.* Olomouc: Vydavatelství Univerzity Palackého. (V tisku.)

Bibliografie k textu:

BOONEN, Ute K. – HARMES, Ingeborg (2013): *Niederländische Sprachwissenschaft. Ein Einführung.* Tübingen: Narr Francke Attempto Verlag GmBH.

CHLUMSKÁ, Lucie (2014): Není korpus jako korpus: Korpusy v kontrastivní lingvistice a translatologii. *Časopis pro moderní filologii*, 96, č. 2, 221–232.

KŘÍŽOVÁ, Kateřina (2015). 'Het Nederlandse woord even en zijn Tsjechische equivalenten in het paralelle corpus InterCorp'. *Brünner Beiträge zur Germanistik und Nordistik,* ročník 29, číslo 2, s. 151–165.

Internetové zdroje

Algemeen Nederlands woordenboek. [cit. 2017-01-03]. Dostupné z: http://anw.inl.nl/

Alpino Treebank Tools. [cit. 2016-11-15]. Dostupné z: http://www.let.rug.nl/~vannoord/alp/Alpino/TreebankTools.html

Boková, E. – Hrnčířová, Z. – Vavřín, M.: *Korpus InterCorp – nizozemština, verze 9 z 9. 9. 2016.* Ústav Českého národního korpusu FF UK, Praha 2016. Dostupný z WWW: http://www.korpus.cz

Corpus Hedendaags Nederlands. [cit. 2016-11-20]. Dostupné z http://chn.inl.nl/

Čermák, F. – Rosen, A. (2012): The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–427.

Český národní korpus. [cit. 2017-01-06]. Dostupné z: https://www.korpus.cz/

Instituut voor de Nederlandse taal. [cit. 2016-10-18]. Dostupné z: http://ivdnt.org/

SoNaR: STEVIN Nederlandstalig Referentiecorpus. [cit. 2017-01-04]. Dostupné z: http://lands.let.ru.nl/projects/SoNaR/description.html

TST-Centrale. [cit. 2016-12-08]. Dostupné z: http://tst-centrale.org/nl/

Wiki Českého národního korpusu. [cit. 2016-11-21]. Dostupné z: http://wiki.korpus.cz/