Exkurze do Ústavu teoretické a komputační lingvistiky FF UK

Datum konání: 31. března 2017

Místo konání: Ústavu teoretické a komputační lingvistiky FF UK, Praha

Počet účastníků: 16

Zpracovali: Hladík Judita

Před přednáškou

Exkurze do Ústavu teoretické a komputační lingvistiky FF UK v Praze (Celetná 13, 1. patro, místnost 21) proběhla v pátek 31. března 2017. Pedagogický doprovod tvořily doc. PhDr. Zdeňka Hladká, Dr., doc. PhDr. Klára Osolsobě, Dr., Mgr. Dana Hlaváčková, Ph.D., a Mgr. Hana Žižková. Z Brna jsme odjížděli vlakem v časných ranních hodinách. Cesta proběhla bez komplikací a nikdo se neztratil. V Praze jsme z Hlavního nádraží cestu na Celetnou absolvovali pěšky. Na Celetné nás velmi nepříjemně překvapil technický a hygienický stav budovy. Považuji za skandální, že členové akademické obce a významné vědecké pracoviště musí pracovat v tak otřesných podmínkách. Myslím si, že ani více než 20 let odkládaná rekonstrukce budovy nemůže být záminkou pro zanedbávání základního úklidu a nejnutnější údržby. Nesnesitelně páchnoucí záchody, na kterých teče pouze vařící voda, zaprášené pavučiny visící ze stropu posluchárny, která nemá fungující kliku – to jsou jenom dva z mnoha příkladů toho více než neutěšeného stavu.

Přednášky:

1. Stručné představení ÚTKL (doc. RNDr. Vladimír Petkevič, CSc.)
Ústav byl založen v roce 1990 profesorem Sgallem. V názvu ústavu obsažené slovo komputační je velmi zřídka používané, znamená počítačový. Ústav navázal na tradici Laboratoře algebraické lingvistiky a Pražského lingvistického kroužku. Podle doc. Petkeviče je jich v ústavu pět a půl, a to doslova, 5 pracovníků a sekretářka na poloviční úvazek. Ústav spolupracuje s Ústavem Českého národního korpusu a s Ústavem formální a aplikované lingvistiky UK.

Ústav se zabývá morfologickou a syntaktickou analýzou češtiny a dalších jazyků. V oblasti korpusové lingvistiky se věnuje morfologickému a syntaktickému zpracování korpusů češtiny, vytváření paralelních korpusů a vytváření žákovských korpusů. V oblasti formálního popisu jazyka se zabývá gramatickými formalismy. Věnuje se také lingvistické teorii, zvláště teorii syntaxe.

2. Morfologické a slovnědruhové značkování (doc. RNDr. Vladimír Petkevič, CSc.)
Přednáška se věnovala jednotlivým fázím značkování – tokenizaci, větné segmentaci, morfologické analýze a morfologické disambiguaci. Syntaktické značkování je spojeno s vybudováním stromových struktur. To slouží k disambiguaci lexikálních významů. Morfologická disambiguace je proces, který vede ke zjednoznačnění – z často několika interpretací je třeba přiřadit pouze jednu. V českých textech je velmi častá systémová i náhodná homonymie, což ztěžuje analýzu. Při této analýze se používají statistické metody na základě učení (trénovací korpus) a lingvistická pravidla.

3. Tvorba českého národního korpusu (podle přednášky RNDr. Hany Skoumalové, Ph.D., přednesl doc. RNDr. Vladimír Petkevič, CSc.)
V přednášce byly prezentovány kroky potřebné pro vytvoření korpusu. Jsou to sběr a kontrola textů, převod do jednotného kódování (ÚČNK), preprocessing (oprava nejčastějších chyb), segmentace a tokenizace textu, morfologická analýza, desambiguace, postprocessing a zveřejnění. Další část přednášky byla věnována složitosti desambiguace v češtině. Po morfologické analýze připadá na jedno slovo průměrně třináct tagů.

4. Frazémy a disambiguace v češtině (RNDr. Milena Hnátková, CSc.)
Používá se program automatického vyhledávání frazémů a ustálených slovních spojení v korpusových datech. V korpusu SYN v4 jsou označkovány frazémy. Provádí se automatická anotace frazémů a ustálených kolokací. Základem slovníku je Slovník české frazeologie a idiomatiky. Nejsou zařazeny kolokace, u kterých byla vysoká chybovost. Na příkladech bylo předvedeno vyhledávání přísloví (která přísloví se používají nejčastěji).

5. Syntaktická analýza jazyka (Mgr. Tomáš Jelínek, Ph.D.)
Syntaktická analýza byla provedena pouze v korpusu SYN2015. Využívá se při ní automatické značkování syntaktických vztahů ve větě (tzv. parsing). To umožňuje jiný pohled na jazyk než morfologické značkování, užitečné je také v rámci automatického překladu. Pro trénování slouží manuálně označená data v Pražském závislostním korpusu.

6. Hledání v paralelním korpusu (Ing. Alexandr Rosen, Ph.D.)
Při vyhledávání je třeba nejprve vybrat korpus český, a až potom korpus cizojazyčný. Cizojazyčných korpusů je možné přiřadit více současně. Nejčastěji se vyhledávají ekvivalenty nějakého slova, vyhledávat lze i víceslovné výrazy. To je výhodné například při překladech z němčiny, pro kterou jsou víceslovné významy typické, ale použít lze jakýkoli jazyk. Vyhledávat je možné nejen nejčastější výrazy, ale i výrazy a kolokace ojedinělé.

7. Korpus textů nerodilých mluvčích češtiny (Ing. Alexandr Rosen, Ph.D.)
Tyto texty vznikly při výuce v jazykových kurzech různého zaměření a jazykové úrovně studujících. Jazyková úroveň jejich autorů se pohybuje v rozmezí A1 – C2. Korpus je zpracován manuálně, částečně je značkován ručně. Vyhledávat v něm lze i podle rodného jazyka autora a dalších jazyků, které ovládá. Korpus slouží k výzkumu v oblasti osvojování si češtiny a její výuky.

Přednášky byly velmi zajímavé. Škoda jen, probíhaly ve velmi rychlém tempu. Paní docentka Osolsobě požádala přednášející, aby nám svoje prezentace poslali. A tak se můžeme těšit, že se z nich naučíme mnoho nového.

Po skončení přednášek jsme se všichni rozešli. Někteří spěchali zpátky na vlak do Brna, jiní zvolili individuální program v Praze. Ze zajímavých přednášek budeme ještě dlouho čerpat. Můžeme jen doufat, že se naši hostitelé pro svoji práci brzy dočkají civilizovaných podmínek a nebudou již dále muset působit v podmínkách, za které se musíme všichni stydět.

Fotografie (autorem je Kryštof Davídek):