Posudek oponenta disertační práce Petra Plecháče

“Atribuce autorství básnických textů”

předkládané v roce 2019 na Ústavu českého národního korpusu

I. Stručná charakteristika práce

Určování autorství je jedna z nejstarších disciplín, které vnesly do literární vědy kvantitativní metody, přičemž od počátku bývaly užívány v literární vědě běžné prostředky, tedy i prostředky versologické. Obdobně v lingvistice a NLP má určování autorství velmi silnou tradici a stylometrie vyvinula mnoho různých sofistikovaných metod, jak daný problém řešit. Je tedy s podivem, že literárněvědná a stylometrická tradice jsou do značné míry mimoběžné. Tato disertace konečně spojuje obě tradice, čímž vznikají nové metody pro určování autorství básnických textů, úspěšnější než předchozí. Tyto metody jsou následně použity pro atribuci několika anglických a českých veršovaných textů.

II. Stručné celkové zhodnocení práce

Hlavními přednostmi posuzované disertace jsou preciznost a jasná koncepce. Autor přesně ví, čeho chce dosáhnout, a na své cestě si dává velký pozor, aby se nedopustil žádných metodologických ani logických chyb. Vzhledem k tomu, že oblast zájmu je plná nástrah a stále v ní nejsou pevně daná pravidla, není to jednoduché, nicméně s překážkami se se ctí vyrovnává.

III. Podrobné zhodnocení práce a jejích jednotlivých aspektů

První a druhá kapitola nás uvádí in medias res a stručně shrnují metody automatické atribuce textů, jak je známe z NLP i literární vědy. Specificky se zabývají zejména těmi metodami, které jsou následně užity ve zbytku práce, přičemž nešetří místo ani energii na to, aby čtenář skutečně dopodrobna a do hloubky pochopil podstatu líčených metod. Plecháčova schopnost vysvětlit poměrně sofistikovaný fenomén jednoduchými slovy je fascinující. Podstatná energie je zde věnována správné operacionalizaci versologických charakteristik tak, aby je bylo lze využít. Drobná výtka: poněkud zbytečně kontroverzní shledávám tvrzení na straně 32 „[v]ersologické rysy lze považovat za kontextově méně závislé než obvyklé rysy užívané ve stylometrii (frekvence slov, znakových a slovních n-gramů)“, které by bylo možné přímo empiricky testovat.

Třetí kapitola empiricky testuje navržené operacionalizace a analyzuje výsledky, případně navrhuje zlepšení v operacionalizaci, takže se na konci dostaneme k metodologii atribuce autorství, která bude použita v kapitole čtvrté. Autor používá standardní evaluační techniky, takže výsledky, ke kterým přichází, jsou důvěryhodné (což bohužel v NLP není samozřejmost). Drobná výtka: poněkud podivný je model vztahu mezi úspěšností rozpoznávání autora a počtem jeho vzorků na straně 49. Proč je v obou případech použit jiný algebraický model? Jaké principy za těmito modely stojí?

Konečně čtvrtá kapitola využívá metod, které vykrystalizovaly ve třetí kapitole, k atribuci autorství několika anglických a českých textů, což můžeme také považovat za určitý způsob evaluace, neboť *the proof of the pudding is in the eating*. I zde kriticky a velmi přehledně představuje techniky použité jeho předchůdci, následně je komparuje s výsledky vlastními.

1. *Struktura argumentace.*

Práce vyniká naprosto jasnou argumentační strukturou, jak jsme ostatně u autora zvyklí z jeho ostatních publikací. Určitým nebezpečím při hodnocení této práce je její stylistická a vypravěčská obratnost, vše plyne příliš hladce, koherentně a bez náznaků problémů. Vzniká tak dojem, že výzkum probíhal snadno, bez negativních výsledků, bez tápání, bez zklamání, bez chyb, a čtenář tak může mít tendenci podcenit množství práce, které za výzkumem stojí.

2. *Formální úroveň práce*

Po formální stránce je práce výborně zpracována, zejména oceňuji přehledné grafy a tabulky. Co se týče bibliografe, spíše než formální úprava je pro mě důležitější, že autor cituje řádně a poctivě. Nicméně i formální úprava bibliografických údajů je přehledná a přesná. Autor se odvážně neřídí ISO690, ale zvykovou normou používanou v korpusové lingvistice, což oceňuji, neboť ISO690 je sice nekontroverzní, nicméně značně nepřehledná. V případě knižního vydání případné formální nedostatky odstraní redaktor a korektor. Ostatně knižní vydání práce vřele doporučuji.

3. *Práce s prameny či s materiálem*

Kladně hodnotím práci s literaturou, která v přehledových částech (například v krátké kapitolce o české stylometrické tradici na straně 31) není pouhým výčtem, naopak dává vše do kontextu a ke všemu přistupuje kriticky.

Na Plecháčově práci s korpusy je vidět, že není pouze jejich uživatelem, ale že už lecjaký korpus kompiloval, a že „ví, co má uvnitř“.

Velmi oceňuji, že se u použitých anotačních nástrojů nespokojí s údaji o úspěšnosti, které u nich udávají jejich tvůrci, ale že si každý nástroj testuje sám. Určitá míra nedůvěry je zde opravdu na místě. Dokonce se rozlišuje mezi úspěšností anotačního nástroje pro prostředek / konec verše, což je mnohem preciznější, než bych čekal, nicméně dává to velmi dobrý smysl.

4. *Vlastní přínos*

Jak bylo zmíněno v úvodu, práce poprvé spojuje literárněvědný a stylometrický přístup k atribuci autorství, tedy dvě paralelní metodologické tradice, přičemž obě jsou svým vlastním způsobem sofistikované a vyžadují hluboké porozumění. Znamenalo to nejen vnořit se do obou diskurzů a heuristicky je vytěžit, ale také vytvořit přemostění mezi nimi, zejména operacionalizovat tradiční versologické charakteristiky tak, aby mohly výhodně sloužit jako *rysy* pro PCA.
Zároveň se autor nespokojuje s vytvořením a otestováním metodologie, ale rovnou ji vyzkouší ve dvou případových studiích.

IV. Dotazy k obhajobě

Na straně 61 se dovídáme, že v angličtině je kombinovaný model méně úspěšný než čistě lexikální. Napadá mě v podstatě banální vysvětlení, že by za tím mohla stát méně úspěšná automatická anotace versologických rysů. Nicméně přivádí mě to k mnohem hlubšímu problému: pokud zvolený algoritmus vykazuje horší výsledky, když mu předložíme větší množství různorodějších rysů (tedy nedokáže ignorovat rysy, které nezlepšují zvolenou metriku), není to známka toho, že PCA není příliš vhodnou metodou a že bychom ji měli nahradit?

Dále bych byl rád, kdyby se debata stočila na obecnou metodiku evaluace použité metody atribuce. Na stejných datech bylo totiž trénováno a následně testováno několik konkurenčních metod atribuce. Tím se ovšem samotný výběr metody stal svým způsobem trénováním. Abychom určili konečnou úspěšnost metody, kterou zvolíme pro následné praktické užití, měli bychom ji nakonec testovat na datech, na kterých jsme netrénovali ani netestovali žádnou z konkurenčních metod v době výběru.
Již v úvodu autor píše, že „[versologické] charakteristiky bývají tradičně považovány za specifický rys autorské, případně alespoň dobové poetiky“. Bylo by tedy možno nalezené metody použít pro atribuci nikoli autora, ale doby vzniku díla?

V. Závěr

Určování autorství literárního díla, ležící na pomezí kvantitativní lingvistiky, automatického zpracování přirozeného jazyka a machine learningu jako takového, statistiky, stylistiky a různých odvětví literární vědy, je pro svou tematickou šíři v dnešní době spíše doménou menších týmů, nikoli jednotlivce. Disertace Petra Plecháče mě však již od začátku překvapila svou precizností a suverenitou, s jakou se pohybuje v celém prostoru.

Předložená disertační práce splňuje požadavky kladené na disertační práci, a proto ji doporučuji k obhajobě a předběžně ji klasifikuji jako *prospěl*.

V Praze, 4. listopadu 2019 Jiří Milička v. r.