Univerzita Palackého v Olomouci

Filozofická fakulta

Katedra historie

Ivana Platilová

Možnosti aplikace počítačových metod v historii

magisterská diplomová práce

Vedoucí práce: Doc. Mgr. Martin Elbel, M.A., Ph.D.

Olomouc 2019

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a uvedla v ní všechny použité zdroje a literaturu.

V Olomouci dne 16.8.2019

Poděkování

Mé poděkování patří doc. Martinu Elbelovi za odborné vedení, trpělivost a ochotu, kterou mi v průběhu zpracování diplomové práce věnoval.

Obsah

Úvod

Digitální revoluce, jejíž průběh v dnešní době s neustále narůstající rychlostí proměňuje celou společnost, klade mimo jiné i nové nároky na vědu. Zatímco pro přírodní vědu je použití matematických a komputačních metod věcí obvyklou a naprosto přirozenou, informatizace a komputerizace v humanitních vědách způsobuje jisté napětí mezi relativně „subjektivním" objektem a použitím „objektivních" metod. Právě toto napětí se snaží překonat interdisciplinární obor, který se označuje jako „digital humanities", což je aplikace digitálních nástrojů a metod způsobem, který je v soulady se specifiky zkoumaného předmětu. Jednou z poddisciplín tohoto široce pojatého oboru je digitální historie.

Zatímco v anglofonním světě se používání počítačových metod rozvíjí přibližně od padesátých let dvacátého století v podobě kvantitativní historie, od devadesátých let pak zejména v podobě digitální historie zaměřené na webové technologie, v kontinentální části Evropy je zájem o toto odvětví menší (objevil se v osmdesátých letech v souvislostí s použitím databází). Co se týče české historie, rozsáhlejší použití počítačových metod pro ni dodnes zůstávají, navzdory aktuálnosti digital humanities, téměř neznámé. Oproti tomu česká jazykověda se naopak může pochlubit relativně dlouhou tradicí matematické a korpusové lingvistiky, což jsou disciplíny, které, byť nepoužívají název digital humanities, s digitálními metodami aktivně pracují.

Cílem této diplomové práce je poskytnout vhled do metodologického problému používání počítačů při práci s historickými prameny na základě analýzy relevantní historiografie, tj. poukázat jak na možnosti využití konkrétních metod, tak i na problémy a meze, které jsou s nimi spojeny. Dalším cílem je pak předvést v praxi fungování především těch počítačových metod, které se nachází na pomezí historie a kvantitativní, počítačové a korpusové lingvistiky na konkrétních historických pramenech.

Práce je rozdělená na dva logické celky, přičemž první, spíše historiografická část sleduje zrod a vývoj disciplíny digital humanities od jejích počátků až po dnešní dobu se zvláštním přihlédnutím k velmi podstatné složce tohoto směru, kterým je digitální historie. První kapitola se věnuje jednak vymezení disciplíny digital humanities jako vědního oboru, hlavním prvkům, jež ho definují a jeho postavení na poli současné vědy,

jednak poté jeho vývoji ve smyslu aplikace komputačních metod v průběhu 20. století. V této kapitole představujeme historický vývoj tohoto směru na pozadí významných projektů své doby (se zvláštním přihlédnutím k pilotnímu projektu Roberta Busy) a reflektujeme jak úspěšné využití nových metod, tak problémy, se kterými se jejich používání ve svých počátcích muselo potýkat.

Druhá kapitola se věnuje samotné oblasti historie. Blíže se zaměřuje na historickou práci prováděnou pomocí kvantitativních metod, přičemž je zde věnován velký prostor různým využitím těchto metod, ať už se jedná o kliometrii, tj. nahlížení na hospodářské dějiny pomocí statistiky, nebo vytváření historických databází či využití geografického informačního systému v oblasti archeologie. Věnujeme se zde také vývoji historiografie s přihlédnutím na rozdíly, jimiž se vývoj použití počítačových metod vyznačuje v Evropě a Americe. Na pozadí takto představeného vývoje disciplíny se poté pokoušíme odpovědět na otázku, jak bude vypadat její budoucnost a co nového může historickému vědění přinést.

Druhá část práce má více praktický charakter. Postupně zde budou na souboru pramenů korespondenční povahy v podobě souboru 303 textů týkajících se působení anglického diplomata alžbětinské doby Williama Herleho v Nizozemsku představeny komputační metody, spadající tradičně spíše do oblasti lingvistiky, s nimiž může historik při svém bádání pracovat. V celkovém pořadí třetí kapitola obsahuje popis databází, frekvencí slov a frekvenčních slovníků spolu s možností zkoumat tematické zaměření textu, čtvrtá kapitola pak představuje využití jazykových korpusů při práci s historickými prameny. V poslední, páté kapitole, je předvedena ukázka metod strojového učení na příkladu určování autorství. Praktická část je psána tímto způsobem, aby mohla nejen sloužit nastínění různých pomezních interdisciplinárních metod, ale i názorným způsobem předvést možný způsob jejich praktického použití a umožnit pochopení základních principů jejich fungování.

Digital humanities

„Je jasné, že v budoucnu se historici budou potýkat spíše s přebytkem pramenů než s jeho nedostatkem. Už teď v posledních několika letech byly digitalizovány miliony knih, a nepochybně v dalších letech se k tomu přidají ještě další. Knihovny mají k dispozici a zpřístupňují miliony naskenovaných obrázků a dokumentů. Jsou digitalizovány miliony novinových stran, a téměř každý den vzniká nějaký nový digitální historický zdroj nepředstavitelné velikosti.“[1]

Co je to digital humanities – vymezení disciplíny

Matthew Kirschenbaum ve své úvodní pasáži ke článku *What is Digital Humanities and What's It Doing in English Departments* (2012) ze sborníku *Debates in the Digital Humanities* vyslovuje tezi, že pokud dnes ještě někdo pokládá otázku, co je to digital humanities, zřejmě se příliš neobtěžoval s hledáním odpovědi. [2] Diskuze ohledně předmětu a rozsahu této disciplíny se vede již dlouhou dobu – přesto ovšem v následujícím příspěvku v témže sborníku Lisa Spiro oponuje tomuto tvrzení a říká, že namísto debat ohledně toho, jaké metody spadají pod digital humanities a jaké ne a o tom, jak můžeme definovat tuto disciplínu na základě těchto metod, bychom se měli spíše zaměřit na vymezení oboru na základě komunity vědců, která se k této disciplíně hlásí.[3] V následujících kapitolách se pokusíme vyhovět oběma těmto přístupům – co je to digital humanities předvedeme jak na základě přehledu používaných metod, tak i vzhledem k vývoji disciplíny se zaměřením na klíčové osobnosti a pracoviště.

Obecně můžeme tvrdit, že směr digital humanities kombinuje předmět zkoumání spadající do oblasti humanitních studií s metodami a přístupy, které jsou tradičně považovány za součást přírodních věd. V nejširším možném pojetí lze říci, že jakékoli použití komputačních metod ve větším nebo menším rozsahu spojených s humanitními studii má za sebou pozadí této disciplíny, nicméně byla by chyba domnívat se, jak poznamenává Kathleen Fitzpatrick, která zároveň varuje před přílišným rozšířením definice směru, že „každý medievalista s vlastní webovou stránkou dělá digital humanities“.[4]

[1] BERRY, D. Introduction: Understanding the Digital Humanities, s.2
[2] KIRSCHEBAUM, M. What is Digital Humanities and What's It Doing in English Departments, s. 3.
[3] SPIRO, L. „This Is Why We Fight“: Defining Digital Humanities, s. 16.
[4] FITZPATRCIK, K. The Humanities, Done Digitally, s.14.

V porovnání s tradičními humanitními vědami, které se zaměřují vždycky jenom na poměrně malý počet pramenů, se digital humanities snaží pracovat naopak s co největším počtem: David Berry uvádí příklad výzkumu románů vydaných v Británii v 19. století. Existuje určitý soubor dobře probádaných textů, přibližně 200 knih, což ale ve skutečnosti dělá méně než 5 % veškerého počtu románů, které byly v 19. století vydány – jejich celkový počet tvoří korpus, který nelze efektivně zpracovat pouze na základě kvalitativních metod.[5]

Tradiční humanitní vědy jsou zaměřené na texty, které zpravidla mívají až na výjimky jasně vymezený začátek a konec, stejně jako časové určení (byť ne vždy pochopitelně známé), nicméně v současné době je potřeba zpracovávat i informace probíhající v reálném čase, např. analýza sociálních médií, nebo v širším smyslu internetu vůbec, což jsou texty, které nejsou ani uzavřené, ani časově omezené, tedy svým způsobem nekonečné.[6] Proto se digital humanities snaží vyvíjet metody, které by namísto omezeného množství pramenů měly pracovat spíše s jejich přebytkem.[7]

Od klasických humanitních věd se tato disciplína nadále odlišuje tím, že pracuje s tzv. komputačním uvažováním. Jednoduše řečeno spočívá komputační uvažování v aplikaci metod programování a informatiky na zkoumaný předmět. Tento přístup má několik základních složek: automatizaci, čímž rozumíme svěření několika drobných opakovatelných úkolů počítači (příkladem může sloužit počítání frekvence slov nebo hledání slov v textu), abstrakci, jež samozřejmě není výlučně záležitostí informační vědy, nicméně počítačové abstrakce mají svou specifickou podobu (např. programovací funkce, která je definovaná nezávisle na konkrétním vstupu, který dostane), dekompozici, která znamená rozdělení komplexních úkolů na co největší počet drobnějších, a algoritmizaci, způsob, jakým lze propojit všechny předchozí složky pomocí formalizované posloupnosti příkazů, složených z logických operátorů a podmínek.[8] Komputační uvažování v praxi znamená umět převést řešený problém do podoby modelů a algoritmů, které jsou zpracovatelné počítačem.[9]

Význam komputačního přístupu se zvyšuje napříč velkým množstvím různých disciplín, včetně uměnovědy, humanitních studií či sociálních věd, což se občas označuje

[5] BERRY, D. Introduction: Understanding the Digital Humanities, s.1-2.
[6] Tamtéž, s.15.
[7] Tamtéž, s.1-2.
[8] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s. 63.
[9] Tamtéž, s. 77.

za komputační obrat. Toto potvrzuje i nárůst zájmu o digital humanities a komputační sociální vědy, který se projevuje ve zvýšení počtu odborných časopisů, konferencí, vědeckých monografií a grantů, jež se těmto disciplínám věnují a o nichž bude řeč v dalších částech práce.[10]

Obor digital humanities je tedy již od svých prapočátků neodmyslitelně spojen se vznikem počítačů. Tento milník znamená přechod od tištěné kultury k digitální – vyžádal si vznik nových nástrojů a digitálních archivů. Toto rozlišení mezi nástroji a archivy, tedy mezi programem a daty, zůstává v disciplíně důležité dodnes. Komputační metody jsou dlouhodobě spojované s novými způsoby ukládání dat jako digitálními archivy a databázemi a rovněž s výpočetní kapacitou umožňující hledání v textech, filtrování, vytváření konkordancí a další manipulace s textem. Tyto nástroje a archivy posloužily jako hlavní přínos digital humanities jakožto vědecké disciplíny, stejně jako samotná komunita vědců zabývajících se touto disciplínou vznikala od dob zakladatele disciplíny Roberta Busy kolem nich.[11]

V dnešní době lze napříč celou univerzitní sférou pozorovat jistou změnu způsobu bádání, a je nasnadě říci, že tato změna je do značné míry způsobena rozvojem digitálních technologií. Čím dál více se výzkum začíná provádět pomocí těchto technologií. Pochopitelně se to některých disciplín týká více než jiných, ale těžko si v současné době dovedeme představit výzkum, který by se neopíral alespoň o nějaké počítačové nástroje – přinejmenším se jedná o elektronické katalogy knihoven či archivů, bibliografické a citační databáze.[12]

Navzdory tomu se šíření samotných výpočetních metod použitých přímo k výzkumu v humanitních vědách místy nesetkává mezi vědci s přílišným nadšením. Zatímco základní počítačová gramotnost v podobě práce s texty, používáním internetu a e-mailu je mezi humanitními vědci všudypřítomná, jejich metodologické a technické schopnosti ve sféře digitalizace výzkumu jsou velice omezené.[13] Aplikace počítačových algoritmů tedy většinou vyžaduje velké interdisciplinární týmy, protože jenom velice omezené množství vědců má dobré znalosti zároveň v oblasti matematiky a zároveň v oblasti humanitních věd. Vznik takových týmů se často označuje jako big humanities,

[10] BERRY, D. Introduction: Understanding the Digital Humanities, s.11.
[11] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s.45.
[12] BERRY, D. Introduction: Understanding the Digital Humanities, s.1
[13] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s.14.

což má znamenat posun od kvalitativního výzkumu prováděného jednotlivci k týmovému kvantitativnímu výzkumu kombinujícímu počítačové metody s metodami humanitních věd, založenému na velkém množství dat (tzv. big data).[14] Rovněž se napříč celou akademickou obcí, a dokonce i mimo ni se zvyšuje zájem o disciplínu, která je označovaná jako data science, což je interdisciplinární přístup hledající zákonitosti a korelace ve velkém množství dat. Aplikace tohoto přístupu v humanitních vědách je nemyslitelná bez přispění digital humanities a přístupů, které se s touto disciplínou pojí.[15]

Digital humanities má zároveň tendenci chápat výstupy akademické práce šířeji než tradiční disciplíny. Zatímco tradiční disciplíny se zaměřují především na monografie a vědecké články, odborníci na digital humanities rádi pracují s takovými nestandardními výstupy jako jsou například databáze nebo vizualizace. Z toho vyplývá, že digital humanities jako disciplína vnímá sebe sama jako outsidera na poli tradiční humanitní vědy – běžný pohled je stále takový, že digitální metody jsou užitečným nástrojem při sběru dat, ale nikoli při interpretaci.[16]

Původně se tento směr nazýval humanities computing nebo computing in the humanities, což do češtiny můžeme volně přeložit jako aplikace výpočetních metod v humanitních vědách. V této době se jednalo pouze o jednotlivé metody, které měly být použity k určitým výpočtům v rámci humanitního výzkumu. Později ale dochází ke změně, kterou vystihuje i změna názvu – s přechodem od digital computing k digital humanities vzniká samostatná disciplína, která má svoje vlastní metody a výzkumné standardy a v níž používání digitálních technologií již není redukováno na úroveň pouhého nástroje.[17]

Jejím původním záměrem bylo přiblížit onu metodu dříve zvanou humanities computing pro humanitní vědce. Disciplína vychází především z projektů, které se zaměřují na aplikaci výpočetních metod na texty. Některé z projektů měly povahu digitálních archivů, které zpřístupnily velké sbírky textu pro akademické účely. Vývoj nástrojů a technických standardů pro práci s takovými projekty vždy fungoval jako základ disciplíny digital humanities.[18]

[14] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s.72.
[15] Tamtéž, s.56.
[16] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s. 46.
[17] BERRY, D. Introduction: Understanding the Digital Humanities, s.3.
[18] FITZPATRCIK, K. The Humanities, Done Digitally, s. 13.

V rámci celkového vývoje disciplíny můžeme mluvit o několika velkých vlnách. První, období humanities computing, jak o něm bude řeč dále, se datuje od vzniku disciplíny během druhé světové války do roku 2001; druhé pak v souvislosti s rozvojem internetu spadá do let 2002 až 2009. V současné době podle Berryho probíhá třetí vlna, jež by měla vytvořit disciplínu, jejímž úkolem bude nejen poskytovat metody pro disciplíny jiné, ale vytvářet samostatné vědění sama o sobě. Ta by také měla reflektovat vliv digitalizace na lidskou kulturu a společnost, tzv. „komputační obrat".[19] Jeffrey Schnapp a Todd Presner naopak nezohledňují první období humanities computing, ale hovoří rovnou o digital humanities, přičemž jejich rozdělení odpovídá Berryho druhé a třetí vlně. První vlna podle nich probíhala na přelomu 90. let a začátku milénia, a zaměřovala se na digitalizační projekty, druhá vlna pak v současné době pracuje především s digitálními médii.[20]

Stephen Ramsay poté dělí digital humanities na dva základní proudy, přičemž první proud vnímá disciplínu jako prostý soubor nástrojů a archivů, což znamená, že není spojená předmětem, ale spíše metodou, která zahrnuje jakékoli používání počítačů ve výzkumu, včetně vytvoření archivu, textové analýzy, map, trojrozměrných vizualizací apod.; druhý proud je poté reprezentován především zkoumáním médií, sociálních sítí a internetu a vychází spíše z mediálních a kulturálních studií.[21] Kathleen Fitzpatrick se ztotožňuje s Ramsayho členěním disciplíny a upozorňuje na jisté napětí mezi oběma proudy – první chce především vytvářet a ten druhý interpretovat. Nadále podle ní existuje kreativní napětí mezi vědci, kteří jsou součástí oboru již dlouhou dobu a mezi nově příchozími, rovněž napětí mezi disciplinaritou a interdisciplinaritou a mezi vytvářením a interpretací.[22]

Tom Scheinfeldt nadále tvrdí, že obor digital humanities má dvě základní větve, jednu literární a jednu historickou, které se původně rozvíjely nezávisle na sobě a začaly splývat až na konci 90. let s rozšířením internetu.[23]

Samo pojmenování digital humanities se poprvé objevilo v roce 2001. V tomto roce byl na Université Laval v Quebecu otevřen studijní obor pod tímto názvem. Původně se diskutovalo o názvu humanities informatics, ale tento název byl zavržen z důvodu

[19] BERRY, D. Introduction: Understanding the Digital Humanities, s.4.
[20] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s. 53-54.
[21] Tamtéž, s.54.
[22] FITZPATRCIK, K. The Humanities, Done Digitally, s.13-14.
[23] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s.42.

příliš technokratického znění. Starý název humanities computing ale použit nebyl, protože pro autory programu evokoval dojem podpůrné disciplíny či nástroje, nikoli samostatného vědního oboru.[24] V současné době čím dál víc roste počet univerzit, které zařazují digital humanities do svých výzkumů nebo do výuky a neustále se zvyšuje i financování disciplíny, a to tempem, které nebylo na konci 90. let vůbec myslitelné.[25]

Ještě před nedávnem se digitální metody používaly především jako podpůrný nástroj spíše než samostatný kritický přístup. Používání počítačů k samotnému vytváření archivů nebo k vyhledávání v textech je v současné době akceptované jako legitimní napříč celou sférou humanitních věd. Jinak je tomu ale v případě vnímání digital humanities jako samostatného vědeckého oboru – pořád přetrvává hranice mezi akademickým světem humanitních věd a technickým světem digitálních nástrojů.[26]

Nicméně, skutečnost, že je dnes digital humanities etablovaná vědecká disciplína, se dá potvrdit tím, že v současné době existuje asociace vědců v tomto oboru, která se jmenuje Alliance of Digital Humanities Organizations (založená 2005). V rámci této asociace se konají každoroční konference pod názvem Digital Humanities. Nakladatelství Blackwell vydalo již tři vydání *Companion to Digital Humanities* (první vydání vyšlo roku 2005), rovněž existuje série monografií s názvem *Topics in Digital Humanities* vydávaná univerzitou v Illinois. Fungují recenzované časopisy, z nichž nejznámější jsou *Digital Humanities Quarterly* nebo *Digital Studies*, konají se i letní školy digital humanities (mezi nejznámější patří letní škola na University of Victoria). Pochopitelně po celém světě existují i pracoviště a ústavy věnované tomuto oboru – už v roce 2012 jejich počet přesahoval stovku. Nadále se konají nejrůznější workshopy, kolokvia a sympozia. V USA v rámci grantové agentury National Endowment for the Humanities existuje pracoviště s názvem Office of Digital Humanities.[27]

Berry tvrdí, že první a druhá vlna digital humanities se rozvíjely v rámci tzv. normální vědy a nebyly tedy v rozporu s tím, co by snad Imre Lakatoš mohl označit jako tvrdé jádro výzkumného programu humanitních věd. Třetí vlna by ale podle něj mohla poukázat na anomálie, které vznikají v rámci tohoto programu, např. periodizace nebo problém výběru textů pro výzkum (viz oněch 95 % románů 19. století, které stále

[24] Tamtéž, s. 50.
[25] KIRSCHEBAUM, M. What is Digital Humanities and What's It Doing in English Departments, s.9.
[26] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s.52.
[27] KIRSCHEBAUM, M. What is Digital Humanities and What's It Doing in English Departments, s. 3-4.

zůstávají nezpracované). Vyzývá tedy k tomu, abychom se v rámci třetí vlny zaměřili na to, jakým způsobem se počítačový kód prolíná s různými aspekty kultury a historie, mimo jiné, jak tento kód ovlivňuje fungování akademické sféry.[28]

Pro tradiční humanitní a sociální vědy je těžké úplně ignorovat probíhající digitalizaci a vznik různých digitálních archivů, metod a nástrojů souvisejících s oním tzv. komputačním obratem. Ne všichni si ale podle Berryho uvědomují, že tyto nástroje a metody do jisté míry ovlivňují podobu vědění. Výsledná podoba těchto archivů apod. je určena strukturou počítačů a povahou informační vědy. Tento proces digitalizace má velký dopad napříč různými disciplínami a vede ke vzniku specifického post-disciplinárního vědění.[29]

Berry mluví i o jisté krizi univerzit, ke které dochází v souvislosti s rozvojem digitální společnosti. Univerzity přestávají fungovat jako hlavní zprostředkovatel vědění, což může způsobovat dezorientaci v situaci informační záplavy.[30] Existuje názor, že i samo vzdělávání na univerzitách prochází proměnou: zatímco dříve univerzity především zprostředkovávaly a předávaly dále odborné vědění, nyní mají za cíl především vychovat informačně gramotného jedince schopného kritického myšlení, který se umí přizpůsobit novým metodám a zorientovat se v obrovském množství informací, které neustále narůstá. Takový odborník by měl být vzhledem ke zvyšující se míře používání dat a informací schopný rychle a efektivně získat, zpracovat a vizualizovat data.[31]

Je tedy možné, že současný vývoj digitální společnosti může znamenat počátek vědecké revoluce v Kuhnovském pojetí, která způsobí jistou unifikaci vědění v tom smyslu, že všechny vědecké disciplíny budou mít stejné nebo velice podobné tvrdé jádro založené na komputačních metodách. Komputační věda se tak může stát základem veškerých věd. Toto by nicméně nemělo být založeno na jakési hegemonii metod současné komputační vědy, nýbrž na jakémsi humanistickém pojetí technologie, které má teprve vzniknout.[32]

[28] BERRY, D. Introduction: Understanding the Digital Humanities, s. 5.
[29] Tamtéž, s.13.
[30] Tamtéž, s.8.
[31] Tamtéž, s. 14-15.
[32] Tamtéž, s. 9.

Dějiny disciplíny

V následujících dvou podkapitolách stručně nastíníme vývoj používání digitálních metod v humanitních vědách obecně od jeho samotného vzniku. Postupně se zaměříme na „otce“ moderní disciplíny digital humanities, Roberta Busu, a především jeho projekt Index Thomisticus. Následně představíme další pracoviště a osobnosti, jež se aplikaci počítačových metod věnovaly po celé 20. století až po vznik moderní disciplíny, jak ji definujeme výše. Podrobnému zkoumání jednoho z jejích vyústění, digitální historii, se budeme následně podrobněji věnovat ve druhé kapitole.

Roberto Busa a Index Thomisticus

U zrodu disciplíny digital humanities stojí především celoživotní úsilí jezuitského kněze Roberta Busy. Busa přednášel na jezuitské koleji v Gallarate v Lombardii, která dnes spadá pod Papežskou univerzitu Gregoriana. Celá tato instituce, existující již od 19. století, se věnovala především spisům Tomáše Akvinského, přičemž Busovým původním záměrem byl zde výzkum použití předložky „v“ a její vztah ke konceptu „vnitřního“ v Tomášově díle.[33]

Samotný nápad projektu Index Thomisticus jakožto korpusu všech Tomášových děl vznikl během druhé světové války, kdy se Busa pokoušel najít stroje, které by pomohly automaticky vypracovat lingvistickou analýzu psaného textu. V roce 1949 zjistil, že takový stroj je schopná poskytnout firma IBM, čímž začal samotný projekt, který dohromady trval přes 60 let a zahrnoval zpracování dvaceti dvou milionů slov ve dvaceti třech jazycích a devíti různých abecedách, přičemž polovina z toho byla latina.[34]

Od roku 1949 tedy běžela první fáze projektu, která se opírala o možnost zpracování těchto dat za pomoci technologie děrných štítků. Zpočátku byla metodologie otestována na čtyřech Tomášových hymnech, pro které byl zpracován seznam slov a konkordancí. Konkordance, jakožto slovo a jeho bezprostřední kontext, např. věta, byly v biblických studiích používány již ve středověku, přičemž zde poprvé nacházíme tento termín užitý pro jiný kontext než zkoumání Bible, stejně jako se poprvé setkáváme s jejich automatickým vytvořením pomocí počítačů. Busa tímto položil začátek trendu, který se

[33] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 1.
[34] BUSA, R. Foreword: Perspectives on the Digital Humanities, s. XV.

v humanities computing držel po celou dobu 50. a 60. let – automatické konkordance a jejich tvoření se staly hlavním cílem velkého množství nejrůznějších projektů.[35]

Mimo konkordance se Busova raná práce skládala z pěti dalších částí, kterými byly abecední seznam slov s jejich frekvencí a přiřazeným lemmatem, seznam slov seřazených podle frekvencí, seznam slovních tvarů s jejich frekvencemi, seřazený podle lemmat, a věcný rejstřík. Lze tedy pozorovat, že už v této fázi práce Busa, byť nepoužíval tyto pojmy, běžně pracoval s koncepty tokenů, typů a lemmat čili s každým jednotlivým výskytem konkrétního slova, unikátních slovních tvarů těchto slov a počet unikátních slov.[36]

Během následujících pěti let zpracoval další dva miliony slov a v roce 1956 mohla začít druhá fáze projektu, která se namísto děrných štítků opírala o novou technologii magnetických pásek, což výrazně zjednodušilo zacházení s daty, jelikož pokud by měl být celý projekt realizován na děrných štítcích, vážil by ve výsledku 500 tun.[37] V Gallarate bylo následně založeno speciální pracoviště Centre per l'Automazione dell'AnaIisi Letteraria (CAAL), které se mělo věnovat i zpracování řeckých a latinských textů obecně. V roce 1960 toto pracoviště již zaměstnávalo třicet lidí.[38]

V průběhu 60. let se počet zaměstnanců zvýšil na šedesát, přičemž již tou dobou zaměstnanci pracovali na přípravě velkého korpusu netýkajícího se jen Tomášových děl, který počítal 15 milionů slov v devíti různých jazycích (latina, řečtina, italština, němčina, angličtina, hebrejština, aramejština, nabatejština a ruština). Texty, se kterými se pracovalo, byly různorodé, jednalo se o spisy z oblasti filozofie stejně jako jaderné fyziky; nejvýznamnějším projektem byla tvorba korpusu 30 000 slov ze souboru Svitků od Mrtvého moře, starověkých hebrejských spisů, objevených v roce 1947 – tyto texty dodnes disponují obrovským historickým, náboženským a lingvistickým významem, protože se jedná o jedny z nejstarších dochovaných rukopisů v hebrejštině.[39] Je zde tedy patrné, že se pracoviště, které se roku 1966 přestěhovalo z Gallarete do Pisy a následně v roce 1969 do Boulderu v Coloradu ve Spojených Státech, věnovalo aktuálním tématům,

[35] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 6.
[36] Tamtéž, s. 2-6.
[37] BUSA, R. Foreword: Perspectives on the Digital Humanities, s. 16.
[38] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 2-3.
[39] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 3.

s nimiž obor humanities computing běžně pracoval, a sice zejména starší texty, především starověké a středověké.[40]

Během 70. let byla práce na Indexu Thomisticu prakticky završena. Výsledný korpus se skládal ze dvou velkých částí: v první bylo obsaženo 118 spisů samotného Tomáše Akvinského a ve druhé pak 61 spisů Tomášových komentátorů.[41] Konkordance zde měly být vyhledávány na základě přiřazení slovních tvarů k lemmatům. Lemmatizace všech jedenácti milionů slov proběhla poloautomatickým způsobem, což znamená, že část těch slov byla lemmatizovaná automaticky a část ručně. V roce 1974 začaly konkordance vycházet v tištěné podobě.[42]

Na přelomu 70. a 80. let byl korpus uložen na 1 800 páskách, každé přibližně 700 metrů dlouhé; jejich souhrnná délka čítala zhruba 1 500 kilometrů. V této fázi projekt obsahoval databázi, jejímž tišteným ekvivalentem by bylo 65 000 stran.[43]

Od roku 1987 se začalo přecházet k třetí fázi, kdy se texty začaly ukládat na CD-rom, kam byly současně implementovány hypertextové odkazy, jež sloužily ke snadnějšímu přechodu mezi různými konkordancemi.[44]

Velký mezník znamenal nejen pro Busův projekt, ale i pro celou oblast disciplíny pochopitelně vznik a rozvoj internetu. Od děrných štítků přes magnetické pásky a následně ukládání na CD-romy vznikla v roce 2005 současná podoba Indexu Thomisticu, jež je jako moderním způsobem zpracovaný korpus veřejně přístupná na internetu z adresy http://www.corpusthomisticum.org/it/index.age. V současné době korpus zabírá přes jeden GB dat.[45]

Další pracoviště a projekty

Vedle monumentálního Busova projektu vznikala zároveň s rozvojem disciplíny, která spolu s technologickým pokrokem získávala stále více na přitažlivosti, další pracoviště zabývající se různorodou aplikací počítačových metod na rozličné vědní obory. V průběhu 20. století můžeme sledovat vývoj metody i předmětu zájmu, stejně

[40] BURTON, D. M. Automated Concordances and Word Indexes: The Early Sixties and the Early Centers, s. 92.
[41] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 3.
[42] HOCKEY, S. The History of Humanities Computing, s. 4.
[43] BUSA, R. Foreword: Perspectives on the Digital Humanities, s. XVII.
[44] BUSA, R. Foreword: Perspectives on the Digital Humanities, s. XVI-XVII; HOCKEY, S. The History of Humanities Computing, s. 4.
[45] Tamtéž.

jako samotné vykrystalování disciplíny, jež původně pod názvem humanities computing nebo computing in the humanities, sloužila pouze k určitým výpočtům v rámci humanitního výzkumu nebo automatizaci činností, které se dříve vykonávaly ručně.[46]

Zpočátku byl důraz kladen na předmět spadající do oblastí především literární vědy, zejména pro zkoumání básnických textů, biblických studií a lingvistiky. Obecně napříč všemi těmito vědami bylo dominujícím rysem v bádání až do konce 60. let vytváření frekvenčních slovníků a konkordancí. Takto vytvořené seznamy byly obvykle samoúčelné, zejména v 50. letech bylo vytváření konkordancí bráno za samostatný cíl – nebyly určeny k lingvistickým účelům či vědecké textové analýze; nebylo zatím zkrátka jejich účelem sloužit ničemu dalšímu.[47] Postupně automaticky zpracované konkordance začaly převažovat nad ručně zpracovanými a kolem poloviny šedesátých let je definitivně vytlačily – poslední konkordance vytvořené bez použití počítačů byly vydány v roce 1965. [48]

Prvním z projektů týkajících se aplikace počítačové metody na předmět studia literární vědy je projekt z počátku 50. let, kdy let Guy Montgomery pracoval na projektu vytvoření seznamu slov z básní anglického básníka Johna Drydena; po Montgomeryho smrti pokračovala v práci Josephine Miles, která jako jedna z prvních obrátila svou pozornost k frekvencím konkrétních slov a zaměřila se na nejfrekventovanějších slova v básních. Výstupem projektu je publikace *Concordance to the Poetical Works of John Dryden*, vydaná roku 1957. [49]

Dalším dílem je práce Harryho Josselsona a Howarda Hyada z roku 1953. Ačkoli je práce nazvána *The Russian Wordcount*, neobsahuje ani slovníky slov používaných konkrétním autorem ani konkordance, nicméně v průběhu projektu byl automaticky korpus ruské beletrie čítající milion slov, z níž bylo do výstupu vybráno 5 230 v ruštině nejpoužívanějších slov. [50]

Od konce 50. do konce 60. let také vznikala série automaticky zpracovaných konkordancí na Cornellově univerzitě. Celkem se jednalo o pět různých vydání, přičemž každé vydání bylo věnované jednomu anglicky píšícímu autorovi. Na přelomu 50. a 60.

[46] BERRY, D. Introduction: Understanding the Digital Humanities, s.3.
[47] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 9.
[48] BERRY, D., FAGERJORD, A. Digital Humanities. Knowledge and Critique in a Digital Age, s. 45.
[49] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 4.
[50] Tamtéž.

let taktéž vytvořil Thomas Sebeok na Indiana University konkordance lidových písní v čeremištině, což je menšinový ugrofinský jazyk v evropské části Ruska. [51]

Mimo literární texty se konkordance vytvářely také v rámci biblických studií. Ve spojených státech se na Wooster College v Ohiu začalo v polovině 60. let pracovat na digitalizaci bible v rámci projektu The Computer Bible. Mimo amerických vědců na projektu pracovali odborníci z různých částí světa jako mimo jiné výše zmíněný Andrew Morton nebo Yehuda Raddaj z Izraele. Napřed byly zpracovány Markovo, Matoušovo a Lukášovo evangelium na základě řeckého překladu bible. Tyto konkordance vyšly pod názvem *Critical Concordance to the Synoptic Gospels* v roce 1969. Další biblické texty, jako například Janovo evangelium, Listy Římanům a Skutky Apoštolů byly zpracovány v následujících letech.[52]

Ačkoli vytváření konkordancí bylo hlavním účelem nově vznikajících projektů, objevily se již v této době i další nástroje pro analýzu textů, s jejichž konceptem pracujeme dodnes. Taktéž v roce 1959 přišel Hans Peter Luhn, významný badatel v oblasti počítačové vědy, mimo jiné s konceptem klíčových slov, jež bylo původně pojímána jako slova, která vystihují předmět textu víc než slova jiná. Na základě jejich vyhledávání pak začaly vznikat systémy pro automatickou anotaci, což mohlo následně vést k automatickému zpracování bibliografie. V současné době je vyhledávání pomocí klíčových slov neodmyslitelnou součástí práce s jakoukoli vědeckou či jinou databází.[53]

V šedesátých letech se následně začaly počítače používat k určování autorství. Samotná metoda určování autorství je poměrně stará, existuje přibližně od poloviny 19. století, nicméně před 60. lety 20. století probíhala tato tzv. stylometrie vždy ručně. Používání počítačů, které byly schopny zpracovávat frekvence slov, tento proces učinilo přesnější a rychlejší. První náznaky sledujeme v roce 1963, kdy skotský kněz Andrew Morton zpochybnil autorství některých listů sv. Pavla na základě automatických výpočtů, které provedl. Nejznámějším případem použití počítače k určování autorství byla ale v roce 1964 vydaná studie zabývající se dvanácti listů federalistů (Listy federalistů je korpus 85 textů týkajících se politických pamfletů či politických esejů), jejichž autor byl neznámý. Na základě použití statistických metod byl jako nejpravděpodobnější autor ve

[51] Tamtéž, s. 8-9.
[52] BURTON, D. M. Automated Concordances and Word Indexes: The Early Sixties and the Early Centers, s. 87.
[53] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 10.

studii následně vyhodnocen James Madison. Tento výzkum se stal natolik známým, že se začaly listy federalistů používat jako vzorová metoda pro určování autorství.[54]

Na základě rozmachu disciplíny se začaly mimo jednotlivé projekty postupně objevovat i samostatná pracoviště zabývající se aplikací digitálních metod. Již v roce 1959 bylo ve městě Besançon na východě Francie založeno pracoviště Laboratoire d'analyse lexicologique pod vedením francouzského lingvisty Bernarda Quemady, které vytvořilo korpus ze 150 děl francouzských autorů ze 16. až 19. století. Automatickému zpracování zde předcházela ruční analýza, poté byly pro mnohé z textů vytvořeny frekvenční slovníky, konkordance a byly také zveřejněny fotokopie, posléze uloženy v archivu. Korpus byl lemmatizován a dokonce se počítalo i se schopností odlišovat homonyma.[55]

60. léta byla ovšem obdobím velkého rozvoje disciplíny obecně; od jejich začátku se mimo jiné za výrazného přispění slavného amerického historika Roberta Fogela, o němž se blíže zmíníme v kapitole věnované kliometrii, začaly digitální metody v humanitních vědách výrazně šířit, a to jak v Americe, tak i Evropě. Začaly se konat vědecké konference, stejně jako vznikala další specializovaná pracoviště, která se zaměřovala především na zpracování frekvenčních slovníků, konkordancí, a dalších nástrojů pro textovou analýzu literárních textů.[56]

První pracovní skupina se objevila ve Velké Británii v Cambridgi pod vedením Roye Wisbeyho a jednalo se o vůbec první počítačové zpracování středověkého textu v němčině. Jednalo se o korpus 28 000 slov z textů Wiener Genesis, iluminovaného manuskriptu pocházejícího pravděpodobně ze Sýrie první poloviny 6. století. Tento korpus uměl pracovat s mírou podobnosti slov pro zkoumání rýmů anebo zobrazovat slova nikoli od začátku, ale od jejich konce. Dále v průběhu 60. let stejné pracoviště vypracovalo korpusy dalších středověkých textů, např. Speculum ecclesiae, tj. sbírka kázání v latině a střední hornoněmčině z 12. století. Na konci 60. let byla podobným způsobem zpracována Píseň o Rolandovi. V roce 1964 Wisbey založil pracoviště Literary and Linguistic Computing Centre, kde se konaly mimo jiné vědecké konference věnované počítačovému zpracování literárních textů, jejichž počet na konci dekády přesáhl sto.

[54] HOCKEY, S. The History of Humanities Computing, s. 5.
[55] BURTON, D. M. Automated Concordances and Word Indexes: The Fifties, s. 10-11.
[56] BURTON, D. M. Automated Concordances and Word Indexes: The Early Sixties and the Early Centers, s. 83.

Zároveň mělo také toto pracoviště také největší sbírku textů v moderní arabštině a střední hornoněmčině.[57]

Dalším průkopníkem humanities computing na britských ostrovech byl v oblasti literatury původně novozélandský vědec Trevor Howard-Hill, který se věnoval počítačovému zpracování díla Williama Shakespeara. Prvním zpracovaným textem byla hra *Něco za něco*, přičemž jeho původním záměrem bylo zkoumat Shakespearův pravopis. Výstupem tohoto projektu byla práce s názvem *Oxford Shakespeare Concordances,* publikovaná mezi lety 1969 a 1973.[58]

Ve Skotsku mezitím v polovině 60. let začal pod vedením lexikografa Adama Aitkena, jedné z hlavních osobností v oblasti výzkumu skotštiny, vznikat archiv starších skotských textů (psaných ve skotské angličtině). Cílem bylo vytvoření slovníku staré skotštiny, který pokrývá období od 12. do 18. století. Práce na tomto díla trvala téměř půl století – korpus byl dokončen až v roce 2003.[59]

Kontinentální Evropa začala záhy nastupovat na vlnu nově vznikajícího trendu. V nizozemském Utrechtu bylo v roce 1960 založeno mechanolingvistické centrum pod vedením původně italského lingvisty Maria Alineie, jehož hlavním cílem se stalo vytvoření reverzního slovníku italštiny, který obsahoval seznamy slov podle jejich koncovek. Alinei pracoval se středověkými italskými texty 10. - 13. století, přičemž při zpracování každé slovo obsahovalo informaci o tom, zda je to poslední slovo v řádku (důležitá vlastnost pro zkoumání rýmů) nebo poslední slovo ve větě (pro lingvistické účely); při zpracování se oddělovala homonyma a byla zde poskytnuta možnost vyhledávat u slov další gramatickou informaci. Později v 70. letech se toto pracoviště zabývalo moderními italskými texty ve spolupráci s Accademia della Crusca, jazykovou akademií věnovanou výzkumu a regulaci spisovné italštiny se sídlem ve Florencii.[60]

Na univerzitě v Leidenu v roce 1967 vznikl ústav pro nizozemskou lexikologii pod vedením Féliciena de Tollenaere, který se zabýval digitalizací a vytvořením frekvenčního slovníku Wulfilovy bible. Hlavním cílem bylo umožnit snadné vyhledávání slovních tvarů, aby mohli lingvisté, kteří se věnují gótštině, upřesnit, zdali je v textu bible

[57] Tamtéž, s. 83-85.
[58] Tamtéž, s. 85-86.
[59] Tamtéž, s. 86.
[60] Tamtéž, s. 89-90.

doložený určitý slovní tvar – v zásadě se tedy jednalo o pomocnou metodu při rekonstrukci gótského jazyka.[61]

V sousední Belgii na univerzitě v Liege vzniklo středisko věnované klasické filologii, především tedy latinským textům. Prvním textem byly *Útěchy* od Senecy, přičemž se zkoumaly konkordance, frekvence a byl vytvořen systém automatické lemmatizace pro latinu. V průběhu 70. let zde bylo zpracováno velké množství latinských autorů – Cicero, Seneca, Caesar a mnoho dalších.[62]

Další zemí, která se v této době výrazně podílela na rozvoji disciplíny, byla Itálie. Mimo Busova projektu v Gallarate a Pise vytvořil pro potřeby statistického zkoumání italské fonetiky na začátku 60. let lingvista Antonio Zampolli v Padově program pro automatické rozdělení italských slov na slabiky. Pracoval i na vytvoření lexikálního archivu italštiny a s tímto účelem byly zpracováváni italsky píšící autoři jako Alesandro Manzonni, Boccacio, Petrarca a další.[63]

Další italské pracoviště vzniklo v Římě pod vedením filosofa a historika Tullia Gregoryho. Jednalo se o projekt s názvem Evropský intelektuální lexikon. Můžeme tvrdit, že to byl jeden z prvních – ne-li první vůbec – projektů, který nebyl věnován ani lingvistice, ani literární vědě: mělo se jednat o dějiny filozofie, přičemž cílem projektu bylo vydat sérii monografií věnovaných konkrétním autorům jako například Francisi Baconovi, Giordanu Brunovi a dalším. Studie měly projednávat vývoj filozofické terminologie doprovázený kvantitativní analýzou autorových textů. Zároveň měl vzniknout automaticky zpracovaný lexikon filozofických pojmů 17. a 18. století.[64]

Projekt lexikonu původně vznikl jako pracovní skupina v rámci filozofické fakulty La Sapienza. Dnes spadá pod Consiglio Nazionale delle Ricerche, italskou akademii věd, a jeho důraz je kladen především na dějiny filozofického a vědeckého myšlení ve vztahu k lexikálním jednotkám na pozadí evropské intelektuální tradice od antiky po novověk. Pracoviště kombinuje digitální a korpusové metody s tradičními filologickými a hermeneutickými přístupy. V poslední době se věnuje rovněž

[61] Tamtéž, s. 88-89.
[62] Tamtéž, s. 90-91.
[63] Tamtéž, s. 92.
[64] Tamtéž, s. 93.

papyrologii, migračním studiím, vzniku neologismu v moderní italštině a zachování kulturního dědictví.[65]

Dalšími zeměmi, kde vznikala podobná pracoviště bylo Švédsko, kde se vědecká pozornost upínala především na raně novověké švédské texty nebo Izrael či Německo. V Československu na konci 60. let vytvořila Jitka Štindlová frekvenční slovník Slezských písní.[66] Ten byl manuálně lemmatizovaný a při samotném automatickém zpracování autoři narazili na spoustu problémů, které se v rámci české korpusové lingvistiky řeší dodnes – například existence dubletních tvarů, hláskové alternace, nářečních tvarů, vlastních jmen, cizích slov atd.[67]

Od 70. let se v oblasti humanities computing začaly konat pravidelné mezinárodní konference, kde byli přítomni vědci z Evropy, Ameriky a Austrálie. Evropskou „stranu" zastupovalo pracoviště v Cambridgi, kde se konference tematicky zaměřovaly na programování stejně jako na lexikografii, analýzu textu a stylistiku. Ve spojených státech se od poloviny dekády konala pravidelná konference International Conference on Computing in the Humanities. Konference se konaly každý rok (vždy jednou v Evropě a jednou v Americe), přičemž, ačkoli se obě dvě tato centra zabývala stejnou disciplínou, britská asociace se zaměřovala spíše na lingvistiku a v jejím rámci pak následně vznikla samostatná disciplína korpusové lingvistiky, kdežto americká se vyznačovala širším záběrem a věnovala se rovněž aplikaci digitálních metod při výuce a při studiu umění a archeologie. Není tudíž náhodou, že i disciplína později nazvaná digital history, jíž se budeme ještě obšírně věnovat, vznikla na americkém kontinentu.[68]

Zajímavou okolností rozvoje v 70. letech bylo postupné pronikání disciplíny do výuky na vysokých školách, především v podobě kurzů zaměřených na počítačové aplikace a software a programování. Již touto dobou probíhaly ohledně schopnosti psát programy debaty – vedly se diskuze, zdali by tato schopnost měla být součástí mentální výbavy všech univerzitních studentů. Jednalo se především o znalost programovacího jazyka SNOBOL, který byl speciálně vyvinut pro práci s texty. Autorka Susan Hockey

[65] Introduction - ILIESI Institute Lessico Intellettuale Europeo e Storia delle Idee. Dostupné z: http://www.iliesi.cnr.it/EN/index.shtml
[66] BURTON, D. M. Automated Concordances and Word Indexes: The Early Sixties and the Early Centers, s. 95.
[67] ŠTINDLOVÁ, J., MACHÁČKOVÁ E. Texty Slezských písní Petra Bezruče prověřovány stroji. In: Slovo a slovesnost, ročník 31, číslo 2, s. 161-166.
[68] HOCKEY, S. The History of Humanities Computing, s. 7-8.

v práci *SNOBOL Programming for the Humanities* tvrdí, že se jedná o nástroj vhodný k použití v humanitních vědách pro textovou analýzu stejně jako pro výzkum v oblasti historie, archeologie a hudby, přičemž pro jeho ovládnutí není potřeba mít žádné předchozí matematické nebo programovací znalosti.[69] Námitky, které padly proti tomuto přesvědčení, spočívaly především v tom, že programování je příliš náročné a zároveň příliš vzdálené humanitnímu výzkumu.[70]

Tyto diskuze se ve větší či menší míře objevují i v současné době, kdy se s rozvojem technologií stále častěji vynořuje otázka programování znovu na povrch. Schopnost práce s digitálními technologiemi, ne-li přímo programovacími jazyky, bývá považována za důležitou i pro humanitní vědce – jak dokazuje i samotná metoda digital humanities, která se v posledních letech stále častěji objevuje na prvních místech plánů vědeckého rozvoje – například Národní výzkumná a inovační strategie pro inteligentní specializaci České republiky 2020 mezi prioritními směry výzkumu jmenuje digital humanities včetně data miningu v humanitních a sociálních vědách, přípravy datových zdrojů pro aplikovaný výzkum ve společenských a humanitních vědách a digitální zpřístupnění kulturního dědictví.[71] Jak ukazují zprávy mezinárodní neziskové organizace European Schoolnet, v posledních letech se všechny státy Evropské Unie snaží zařadit výuku programování do studijních plánů základních a středních škol.[72]

Dalším trendem 70. let byl jiný směr počítačových aplikací, který se s postupem času ukázal být jedním z klíčových, a sice tvoření digitálních archivů. V roce 1971 Michael Hart, pracovník počítačové laboratoře na univerzitě v Illinois, převedl do počítačové podoby Deklaraci nezávislosti, což se považuje za počátek první digitální knihovny Project Gutenberg. Záměrem projektu bylo zpřístupnění co největšího počtu knih pro co největší počet uživatelů v co nejsnadněji přístupných formátech. Gutenberg existuje dodnes a v současné době tato digitální knihovna obsahuje 59 000 knih, jejichž počet se pořád zvyšuje.[73]

Mezi dalšími archivy můžeme zmínit poté například *Oxford Text Archive* založený v roce 1976, jehož záměrem bylo uchování zpracovaných textů poté, co byly

[69] HOCKEY, S. SNOBOL Programming for the Humanities, s. 1.
[70] HOCKEY, S. The History of Humanities Computing, s. 9.
[71] Národní výzkumná a inovační strategie pro inteligentní specializaci České republiky 2020, s. 293.
[72] Computing Our Future: Computer Programming and Coding. Priorities, School Curricula and Initiatives Across Europe, s. 17-27.
[73] Project Gutenberg. Dostupné z: http://www.gutenberg.org/wiki/Main_Page

dokončeny výzkumy a zpřístupnění těchto textů jiným badatelům. Nadále pak například i *Theasaurus Linguae Graecae* na kalifornské univerzitě – jednalo se o velkou databázi textů v klasické řečtině, která zahrnovala všechny autory od Homéra po zhruba rok 1600 našeho letopočtu.[74]

Digitální archivy pomohly vyřešit krizi, kterou disciplína procházela v 80. letech. Odklon od digitálních metod byl způsoben jednak určitou vyčerpaností témat, jimž se vědci mohli věnovat, a jednak tím, že se od této metody definitivně odtrhla korpusová a počítačová lingvistika, jež byly v humanities computing hlavní proudy. Nový směr výzkumu přinesl poté právě přechod od konkordancí a textové kritiky k vytvoření digitálních archivů. Archivy nebyly dlouhou dobu považovány za důležité, neboť neexistoval způsob, jakým by se dal zajistit přístup k datům bez fyzické přítomnosti uživatele na pracovišti, tato situace se ovšem výrazně proměnila s nástupem internetu – nově se digitálním archivům mohl dostat kdokoli, aniž by musel být osobně přítomen v místě, kde se archiv nacházel.

Vznik internetu znamenal revoluci i pro zveřejnění výsledků vědecké práce, jelikož badatelé přestali být omezováni formátem tištěných knih. Výsledky projektů mohly být nyní zveřejněny okamžitě a bylo velice jednoduché je aktualizovat. Pro některé obory, mezi jinými i historii, to rovněž znamenalo změnu přístupu k výzkumu, jelikož před vznikem digitálních archivních katalogů bylo pro historiky mnohem větším problémem zjištění toho, zda hledaný pramen existuje a pokud ano, v jakém archivu je uložen, než samotné zpřístupnění pramene. Před vznikem digitálních archivních katalogů byli historici často odkázáni k tomu, že se museli spolehnout na neformální zdroje jako například informace od kolegů.[75] Další revoluci pak mimo jiné způsobila možnost díky technologickému pokroku pracovat nejenom s texty, ale i s obrazem, audiozáznamy a videozáznamy.[76]

Důležité pole působnosti, kde se v 90. letech disciplína dále etablovala je oblast univerzitní výuky. Začala se vyučovat jako samostatný obor, mimo jiné na King's College London, McMaster University in Canada, University of Virginia a University of Alberta. Dvě poslední jmenované instituce tuto tradici zachovaly a dodnes figurují jako hlavní

[74] HOCKEY, S. The History of Humanities Computing, s. 8-9.
[75] BILANSKY, A. Search, Reading and the Rise of Database, s. 3-4.
[76] HOCKEY, S. The History of Humanities Computing, s. 13-14.

centra vývoje disciplíny, mimo jiné i v oblasti digitální historie.[77] Právě na univerzitě v Albertě vznikl unikátní Orlando Project, který je průlomový co do volby předmětu zájmu – spojuje se zde používání digitálních metod s metodologickým přístupem, který je mimo humanitní vědy nepředstavitelný, tedy s feministickou historií.[78] Projekt je mimo jiné specifický i v tom, že se neomezuje jen na samotné texty, ale zpřístupňuje rovněž informace i o autorkách těchto textů spolu s historickým (politickým, sociálním, ekonomickým…) kontextem jejich vzniku.[79]

Jak jsme tedy názorně předvedli na tomto stručném historickém vývoji, procházela disciplína v průběhu 20. století proměnou jak co do metodologických postupů, tak do předmětu studia. Zatímco na počátku stálo humanities computing v pozici nástroje užívaného především pro literární vědu a lingvistiku, postupně se rozšířilo i na oblast dalších humanitních věd, přičemž jednu ze dvou v současnosti nejdůležitějších větví tohoto směru, digitální historii, více přiblížíme v následující kapitole.

[77] Tamtéž, s. 16.
[78] Tamtéž, s. 14-15.
[79] The Orlando Project. Feminist Literary History and Digital Humanities. Dostupné z: http://www.artsrn.ualberta.ca/orlando/

Digital history

Definice disciplíny a vymezení základních pojmů

Celkově se o disciplíně digital humanities dá říci, že obsahuje dvě základní složky – práci s databázemi a práci s texty. Databáze předpokládají tabulky obsahující strukturovaná data s položkami menšího rozsahu, texty se potom chápou naopak jako jeden velký soubor, který je přístupný pro vytváření konkordancí, frekvenčních slovníků, stylometrických analýz a dalších. Strukturování databází probíhá pomocí uložení konkrétních dat do konkrétních položek, strukturování textu poté za pomoci značkování a vytváření metainformací. Dá se říct, že v té části digital humanities, která je zaměřená na historii, výrazně převládá databázová složka, zatímco v oblasti zaměřené na literární vědu složka textová.[80]

Ačkoli jak historie, tak i literární věda používají textový materiál jako předmět zkoumání, jejich přístup a metodologie se výrazně liší, což se promítá i do realizace počítačových aplikací. Literární věda je zaměřená na kritickou analýzu samotných textů a na jejich porovnání s jinými texty. Tyto úkoly si vyžádaly vznik právě takových nástrojů jako jsou ony frekvenční slovníky, konkordance a tematická koncentrace, jak jsme se o nich zmiňovali či ještě budeme zmiňovat. Historie je sice také primárně založena na psaných pramenech, ale na rozdíl od literární vědy ji nezajímají ani tak samotné texty jako takové, ale spíše historická realita, která za nimi leží. Tradičně se zaměřuje na faktické okolnosti textů, např. vlastní jména, data narození nebo úmrtí, obchodní záznamy apod. V tomto pojetí text ztrácí do jisté míry význam jakožto text a funguje spíše jako seznam informací.[81]

Často se může stát, že historie a literární věda zkoumají tentýž text, ale z jiných hledisek. Jako příklad může sloužit případ korpusu korespondence Samuela Pepyse, anglického účetního 17. století, který sice vznikl původně pro historické účely, ale později začal být zkoumán i literárními vědci, a dokonce se občas objevuje ve studijních programech filologie. Historici se v tomto textu snažili najít a označit jména osob, místní názvy a data, a proto pro jejich účely bylo vhodnější vytvořit databázi pro snadné vyhledávání těchto údajů, která obsahovala odkazy na původní text. Literární věda se poté

[80] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 36.
[81] Tamtéž, s. 37.

naopak zaměřila na počítání slov a výzkum frází a jazykových konstrukcí stejně jako na stylometrickou analýzu, především kvůli specifickému autorovu stylu, který kombinuje anglicky psaný text s fragmenty ve španělštině, francouzštině a italštině.[82]

Dříve se aplikace počítačových metod v historii v návaznosti na humanities computing označovala jako history and computing, popřípadě jenom history computing nebo historical computing. Dalším souvisejícím pojmem je kliometrie neboli kvantitativní historie, která označuje směr historiografie založený na aplikaci statistických metod. V současné době se ale spíše používají nová označení, která mají zvýraznit posun směrem od staršího ryze aplikovaného a podřízeného postavení počítačových metod k novým přístupům. V evropském kontextu se často objevuje pojem historická informatika (historical information science), zatímco na druhé straně oceánu se setkáváme téměř výlučně s pojmem digital history. Nejedná se pouze o označení stejné věci různými názvy, jelikož evropská a americká tradice aplikace počítačových metod v historii se liší i v metodologickém zaměření: první klade důraz na vytváření a používání databází, druhá na webové technologie.[83]

Lawrence McCrank definuje historickou informační vědu jako interdisciplinární obor, který spojuje předmět historie, kvantitativní metodologii z lingvistiky a sociálních věd s výpočetní a informační vědou a technologií, a zaměřuje se na historické informační zdroje, struktury a způsoby komunikace. Rovněž i Onno Boostra označuje historickou informatiku za „nový interdisciplinární obor, který pracuje s pragmatickými a konceptuálními problémy, spojenými s používáním informačních a komunikačních technologií při výuce, výzkumu a popularizaci historie“.[84]

Zdůrazňuje se, že historická informační věda není ani historií, ani výpočetní vědou, nýbrž samostatný oborem se svou vlastní metodologií, jehož předmětem je historická informace a způsoby, jak tuto informaci vytvářet, získávat, editovat, analyzovat a reprezentovat za pomocí informačních technologií.[85]

Někteří odborníci, například Charles Harvey, pak upozorňují, že aplikace počítačů v historii by měla vést k vytvoření modelů a reprezentací minulosti, a tudíž nestačí definovat tento obor pouze jako poddisciplínu počítačové aplikace. Ačkoli databáze

[82] Tamtéž, s. 38-39.
[83] Tamtéž, s. 18.
[84] Tamtéž.
[85] Tamtéž, s. 10.

mohou být vynikajícím nástrojem, samy o sobě nejsou specificky historické, jedná se pouze o obecné pomůcky. Můžeme tedy mluvit o aplikaci počítačů v historii jenom v případě, pokud se jedná o výrazný přínos pro historické bádání.[86]

Digitální historie jako taková se objevila v 90. letech minulého století a souvisela především se vznikem internetu. Již o samotných počátků používání tohoto nového média začaly ve velkém vznikat nové počítačové nástroje, které výrazným způsobem proměnily povahu prezentování výsledků historické práce či její výuky – ať už se jedná o prosté přenášení těžiště předávání informací z tištěné podoby na web v podobě přístupu ke katalogům knihoven a archivů či odborným časopisům (zmiňme mezi jinými nástroji například JSTOR) nebo používání nových metod a prezentování výsledků výzkumu na webových stránkách.[87]

Tento směr můžeme široce pojmout jako přístup k výzkumu a prezentování minulosti, který pracuje s novými technologiemi využívající počítačové metody, internet a softwarové systémy. Seefeldt a Thomas rozlišují mezi dvěma různými přístupy – jedním z nich jsou projekty digitalizace, kde dochází k digitalizování velkého množství pramenů a druhým samotný badatelský počin v oblasti digitální historie, který je většinou charakterizován úžeji vymezeným výběrem pramenů a materiálů soustředěných kolem konkrétní historiografické otázky.[88]

Jeden ze zakladatelů disciplíny Roy Rosenzweig poté definuje digitální historii jako „přístup ke zkoumání a reprezentaci minulosti, který využívá nových komunikačních technologií, jako jsou počítače a web. Vychází ze základních rysů digitální sféry jako jsou databáze, hypertextualizace a sítě k vytváření a sdílení historických znalostí.“[89]

Vzhledem k relativní novosti digitálních metod (minimálně v porovnání s tradičními médii) a k rychlosti, s jakou se technologie vyvíjí, měli historici dosud možnost pouze začít prozkoumávat, jak bude historie vypadat v podobě digitálního média. Ukázalo se, že nové možnosti značně proměnily výzkumné techniky, ale otázkou zůstává, zdali se změní i historie jakožto věda.[90]

[86] Tamtéž, s. 18.
[87] SEEFELDT, D., THOMAS, W.G. What is Digital History? In: Perspectives in History, 2009, č. 5.
[88] Tamtéž.
[89] Tamtéž.
[90] Tamtéž.

Zastánci digital history jsou přesvědčeni, že ano. Například právě Seefeld a Thomas spatřují definující rys žánru digitální historie v možnosti čtenářů daný předmět zkoumat a vytvářet pomocí nových nástrojů vlastní interpretace. Místo klasické studie je jim předkládána spíše sada interpretačních prvků, pomocí které mohou získat nový náhled na daný problém.[91]

Tvrdí se, že klíčové digitálně historické projekty jako je *The Valley of the Shadow* nebo *Los Angeles and the Problem of Urban Historical Knowledge* ustanovily nové modely historického výzkumu, které jednak „demokratizují" minulost a jednak se pokoušejí aplikovat alternativní historické, teoretické a metodologické přístupy; vycházely vždy z hlavního tématu či otázky jako je sociální historie americké občanské války nebo způsob, jakým lze mapovat požadavky na znalosti dynamické metropole.[92]

Mnohé z projektů opírající se o webové stránky tedy slouží především k oné „demokratizaci" historie, snadnějšímu přístupu k informacím o minulosti a mají za cíl oslovit širší publikum, především učitelů a studentů. Mnohdy jsou na těchto webech zpřístupněné prameny v elektronické podobě, ať už povahy textové nebo ve formě fotografií.[93]

Počátky history and computing a období kliometrie

Mluvíme-li o digitální historii nebo historické informační vědě, je na místě znovu připomenout, že tato disciplína jako taková vznikla později, přičemž vývoj, který tomuto směru předcházel, byl v Americe odlišný od toho v Evropě. Ve Spojených státech se zájem o aplikaci počítačových metod v historii objevil hned po druhé světové válce a dosáhl svého vrcholu v 60. letech, a to především v oblasti sociální a hospodářské historie. Tento směr, označovaný jako kliometrie, si vypůjčil metody a nástroje ze sociologie a ekonomie a zaměřoval se především na výpočetní a kvantitativní aspekt bádání. K nejznámějším autorům patří Stephan Thernstrom, Robert Fogel či Stanley Engerman.[94]

Vývoj probíhal jinak v západní Evropě, kde sociální a ekonomické historie nejevila příliš velký zájem o výpočetní metody. Badatelé náležící k nejznámější evropské

[91] Tamtéž.
[92] Tamtéž.
[93] Tamtéž.
[94] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 25.

škole, Annales, sice zkoumali sociální a hospodářské dějiny, nicméně přistupovali k bádání povětšinou bez použití počítačů. Jediná skupina evropských historiků, která používala počítač pro výzkum, byli badatelé v oblasti historické demografie.[95]

V Americe po druhé světové válce přišla vlna zájmu o historii jako o sociální dějiny a základními prameny, které se zkoumaly, byly seznamy voličů, městské vyhlášky, daňové listiny a záznamy sčítání lidu. Ačkoli se jedná o různé druhy pramenů, všechny mají jeden společný rys, kterým je skutečnost, že se jedná o texty, které jsou dobře strukturované a jsou tedy poměrně dobře převoditelné do počítačové podoby. Tohle je jeden z faktorů, které umožnily rychlý vývoj počítačových metod v historii na americkém kontinentu. Docházelo zde ke sblížení sociologie, ekonomie, politologie a historie.[96]

Ojedinělé použití počítačů v historickém bádání nacházíme v Americe už na konci 40. let, např. v roce 1949 Frank a Harriet Owsley vydali studii věnovanou kvantitativní analýze sociální struktury jižanských států před občanskou válkou. Byly zpracovány tisíce dokumentů týkajících se sčítání lidu z několika hrabství ve čtyř různých státech a byla připravená statistika věnovaná počtu obyvatel, počtu domácností, počtu otrokářů atd. Výsledky práce naznačovaly, že jih nebyl ve skutečnosti politicky ovládaný především velkými plantážníky, nýbrž spíše drobnými farmáři.[97]

Tímto začala první fáze použití počítačů v historii, kdy byly tyto metody používané ještě velmi zřídka.[98] Na konci 50. let se počítače začaly aktivně používat pro výzkum na univerzitách. Jednou z prvních studií tohoto druhu byla práce Stephana Thernstroma s názvem *Poverty and Progress: Social Mobility in the 19th Century*.[99]

Další fáze začala v 60. letech s rozvojem kliometrie a vyznačovala se zvýšeným zájmem o sociální a hospodářské dějiny. Preferovanými tématy byly migrace, politická příslušnost, urbanizace a asimilace přistěhovalců. Příznivci nových metod sdružovala organizace pod názvem *Inter-University Consortium for Political and Social Research* (ICPSR) založená v roce 1962 za podpory americké historické asociace a americké politologické asociace. Politologové díky počítačům získali možnost zkoumat chování

[95] Tamtéž, s. 26.
[96] GREENSTEIN, D. Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States, s. 351.
[97] THOMAS, W.G. Computing and the Historical Imagination, s. 59.
[98] Tamtéž.
[99] GREENSTEIN, D. Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States, s. 351-352.

voličů v minulých dobách, sociální historici nástroj pro výzkum každodennosti a hospodářští historici možnost vyvíjet složité ekonomické modely s velkým počtem proměnných.[100]

Problémem této doby byla ovšem velká finanční náročnost používání počítačů, objevovaly se i skeptické hlasy, které tvrdily, že soudobé počítače jednoduše nemají dostatečnou výpočetní kapacitu na to, aby mohli uspokojivě zpracovat rozsáhlá historická data.[101]

Přesto se na přelomu 60. a 70. let zdálo, že historiografie, především americká, stojí na pokraji výrazné změny. Rodila se zde nová historie ovlivněná především sociologií, sociologickou teorií a metodologií, která opouštěla tradiční způsob historického bádání a jako první začala systematicky používat počítání a statistiky. Pro některé z historiků to znamenalo nový náhled na historické otázky stejně jako na přístup k práci s prameny, jiní v tom naopak spatřovali možné omezení tvořivé práce historika redukované na jednoduchý počítačový kód. V podstatě se tedy v souvislosti s rozvojem kvantitativní historie začaly ozývat kritické hlasy, například Arthur Schlesinger Jr., jeden z nejznámějších amerických historiků své doby, prohlásil, že skoro všechny důležité historické otázky jsou důležité právě z toho důvodu, že nejsou kvantifikovatelné.[102]

Na začátku 70. let kolem 150 amerických historiků používalo počítače pro svoje výzkumy. Později v průběhu dekády se jejich počet zvýšil natolik, že to podle Joela Silbeyho nebylo možné spočítat. Na přelomu 60. a 70. se začaly konat letní školy věnované tomuto tématu a v rámci americké historické asociace byl založen výbor pro kvantitativní data v historii. Během těchto let také vzniklo několik specializovaných časopisů zaměřených na aplikaci digitálních metod – nejznámějšími jsou Computers and the Humanities (1966), Historical Methods Newsletter (1967), Journal of Interdisciplinary History (1970) nebo Social Science History (1976).[103]

Už ke konci 60. let se začali objevovat první vážní odpůrci, přičemž výtky byly povětšinou následujícího charakteru: používání počítačů zkresluje historické bádání, protože se snaží komplexní historická fakta vtěsnat do pevných matic a čísel, čímž

[100] THOMAS, W.G. Computing and the Historical Imagination, s. 59.
[101] Tamtéž, s. 59-60.
[102] Tamtéž, s. 56.
[103] GREENSTEIN, D. Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States, s. 352.

dochází k přílišnému a zbytečnému zjednodušení. Nejvíce byl kritizován postup kvantitativních historiků v případech, kdy musela být nejednoznačná informace pocházející z pramene jednoznačně zakódována v počítači – historický pramen jako takový je podle nich příliš komplexní, aby se dal redukovat na jednoduchou posloupnost čísel. Ostré kritiky se dostalo například Charlesovi Tillymu za jeho práci o občanských spolcích v západní Evropě 18. a 19. století. Tilly se pokusil o počítačovou analýzu na základě zpracování literárních pramenů. Jeho nejhlasitějším kritikem byl Richard Cobb, podle něhož byly prameny, ze kterých Tilly vycházel, příliš nejednoznačné, aby se z nich daly vyvozovat nějaké závěry. Zastánci kvantitativní historie se naopak obhajovali tím, že přísné požadavky počítačů nutí historiky, aby vyjadřovali svoje závěry více explicitně a aby se vyhýbali vágnostem a dalším nejednoznačnostem. Robert Fogel k tomuto tématu řekl, že neúplnost a nejednoznačnost historických dat pouze vyžaduje vývoj více propracovaných a sofistikovaných statistických metod.[104]

Největší kontroverze vyvolala právě jeho publikace, již vydal ve spolupráci se Stanleym Engermanem, s názvem *Time on the Cross: The Economics of American Negro Slavery* (1974). Tato kniha, kterou můžeme označit za vynikající ukázku aplikace kvantitativních metod v historii (ve výsledku Fogelovi kvantitativní historie vynesla Nobelovu cenu za ekonomii v roce 1993), se skládala ze dvou svazků, z nichž první byl věnován pouze popisu použitých metod a dat. Hned v úvodu se oba autoři přiznávají, že se může jednat o poněkud znepokojující čtení – jak napovídá název, práce se orientuje především na hospodářský dopad a ekonomické fungování otroctví v jižanské společnosti bez přihlédnutí k jakémukoli morálnímu hledisku a hodnocení tohoto jevu.[105]

Tato práce se dle očekávání setkala s obrovskou vlnou kritiky, za prvé, protože Fogelova zjištění byla založena pouze na kvantitativní analýze a věnovala se pouze hospodářským otázkám (jinými slovy, ignorovala kvalitativní analýzu a sociálně-politický kontext otroctví) a za druhé i čistě kvůli tomu, že samotný předmět práce se týká nejkontroverznějšího tématu amerických dějin. V některých případech čelil Fogel kritice poměrně zdrženlivé, například Comer Wann Woodward, renomovaný historik amerického jihu, vyčítal publikaci, že napadá veškerou tradiční interpretaci otroctví, ale

[104] Tamtéž, s. 352-353.
[105] THOMAS, W.G. Computing and the Historical Imagination, s. 56.

na druhou stranu poukázal na to, že by byla velká škoda, kdyby kvůli kontroverzím kolem této publikace začala být zatracována samotná metoda.[106]

Nejradikálnější kritika naopak přišla ze strany Herberta Gutmana v publikaci *Slavery and the Numbers Game*. Gutman zcela ponechal stranou zhodnocení statistických metod a zaměřil se pouze na závěry Fogelovy práce a na kvalitu použitých pramenů stejně jako na způsoby jejich použití. Jednalo se podle něj o výrazně nepovedenou práci, která vykreslovala otroctví ve zlehčujícím, a dokonce snad příznivém světle – především kvůli závěrům tvrdícím, že otroci se jednak ekonomicky vyplatili a jednak jejich postavení nebylo tak špatné, jak se tradičně předpokládá. Autoři podle něj příliš zakládají na předpokladu, že otrok byl zcela pod kontrolou majitele, čímž zanedbávají jeho vlastní osobnost a pohnutky; Gutman se také výrazně vymezuje vůči skutečnosti toho, že autoři často pouze uvádějí některé statistiky, např. počet zbičování ročně, aniž by zhodnotili skutečný sociální kontext, v tomto případě bičování jako výchovný nástroj.[107]

Ke konci 70. let boom kvantitativní historie vyvolal poměrně velkou vlnu odporu, protože se kvantitativním historikům začalo vyčítat, že používají statistiku samoúčelně a že jejich práce jsou příliš popisné, nedostatečně teoretické a nejsou schopny odpovídat na otázky tradiční historiografie. Nadále byl kliometrii často vyčítaný příliš mechanistický přístup a zanedbávání individuí. Postupné narůstání těchto útoků vedlo k uzavření komunity kvantitativních historiků a ztrátě dialogu uvnitř historické disciplíny, což překvapivě přišlo v době, kdy vznikaly první osobní počítače a počítačové metody tedy začínaly být výrazně uživatelsky přístupnější. Příznivci kliometrie nadále pokračovali ve vyjetých kolejích, zatímco zbytek historiků o počítačové metody přestal jevit jakýkoli zájem.[108]

Po vydání *Time on the Cross* se kliometrie v Americe octla v defenzivě, komunita kvantitativních historiků se začala uzavírat a používání počítačových metod začalo upadat. Později byl odpor vůči kvantitativním metodám tak velký, že se všechny digitálně-historické projekty, které zde vznikaly v 90. letech, točily kolem digitalizace zdrojů a používání jednoduchých počítačových nástrojů, jako jsou třeba bibliografické programy, zatímco komplikovanější nástroje a statistika byly zcela ponechány stranou.[109]

[106] Tamtéž.
[107] Tamtéž, s. 57-58.
[108] Tamtéž, s. 355.
[109] Tamtéž, s. 60-61.

Na druhé straně oceánu, v Británii, se podobný odpor vůči kvantitativním metodám v historii nikdy nevyvinul, např. britský historik Charles Harvey poukazuje na to, že britská historiografie vždy měla silný empirický experimentální základ a nezaobírala se příliš teoretickými problémy. Britští historici byli zároveň trénovaní k použití speciálních dovedností a aplikovaných nástrojů, což vytvořilo příznivější prostředí pro použití kvantitativních metod. I zde se ovšem ozývaly kritické hlasy, které spojovaly počítačové technologie s postmodernismem a shledávaly v aplikaci těchto metod rezignaci na objektivitu a historickou pravdu.[110]

Před 80. lety byla aplikace počítačových metod v evropské historiografii velice vzácná, výjimku tvořila pouze britská historická demografie, např. Williams, Speck a Gray za pomoci počítačů připravili analýzu volebních seznamů z 18. století. Jednalo se ale o především o ojedinělé badatele, kteří nezaložili žádné akademické instituce ani časopisy. Jediným střediskem věnovaným aplikaci kvantitativních metod v historii bylo Cambridge Group for Population Studies pod vedením Petera Lasletta založené v roce 1964.[111]

Ve Francii, ačkoli na začátku 70. let škola Annales vydala dvě čísla svého časopisu věnovaná kvantitativním metodám, se používání počítačů příliš neuchytilo. Důvodů, proč se tak nestalo, bylo více, mezi jinými jmenujme skutečnost toho, že na francouzských univerzitách nebyly počítače tak snadno přístupné jako v Americe – přístup k nim měli pouze historici z akademie věd. Dalším důvodem byl fakt, že počítačové nástroje nebyly tou dobou schopny efektivně pracovat s málo strukturovanými daty, s nimiž se historik běžně setkává při práci především se středověkými a raně novověkými prameny.[112]

Výjimkou bylo Německo, kde skupina historiků a sociologů založila v roce 1975 Quantum Group, jež se měla věnovat kvantitativní historii a počítačovým metodám. O několik let později začal být v rámci této instituci vydáván časopis Quantum Information, nyní známý jako Historical Social Research, německy Historische Sozialforschung. Nejednalo se nicméně ani zde o příliš vlivnou instituci.[113] Výrazně větší vliv měli

[110] Tamtéž, s. 61.
[111] GREENSTEIN, D. Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States, s. 355-356.
[112] Tamtéž, s. 356.
[113] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 26.

historici soustřední kolem Ústavu Maxe Plancka, kteří se zaměřovali na počítačové zpracování regionálního vývoje protoindustrializace. Podobná pracoviště začala následně vznikat na konci 70. let i v severských zemích, Británii i ve Francii.[114]

Zajímavým počinem je také práce skupiny kolem Josefa Smetse pod názvem *Languedocienne society (1750-1850)*. Jednalo se o pokus o počítačové zpracování sociální historie v Braudelovém duchu *longue durée*. Pro potřeby projektu bylo digitalizováno a zpracováno velké množství pramenů umožňujících rekonstrukci každodennosti obyvatel Languedocu, především záznamy z farností, soudní záznamy, záznamy sčítání lidu, smlouvy, závěti a další právní dokumenty. Povaha dat, s nimiž Smets pracoval, se výrazně lišila od toho, s čím pracovali američtí kvantitativní historici – nejednalo se o jednoduché jasně strukturované tabulky, nýbrž o mnohem méně jednoznačná textová data. Jeho přístup, v němž se podoba databáze odvíjí od podoby zkoumaných pramenů, v dalších letech převzalo mnoho jiných evropských historiků.[115]

Právě na přelomu 70. a 80. let, kdy používání počítačových metod začalo upadat v Americe, v Evropě se naopak začínalo dostávat do módy. Zatímco ve Spojených státech byla kvantitativní historie založena především na zmiňovaných strukturovaných datech a statistikách, v Evropě převládalo používání relačních databází, zpracování textu a obrazu a práce s nejednoznačnými, nestrukturovanými daty z období středověku a raného novověku.[116]

Od začátku 80. let se začali scházet Evropští historici se zájmem o používání počítačových metod, ze začátku nepravidelně. První takové setkání se uskutečnilo v anglickém Hullu v roce 1983. Diskutovaly se zde především otázky návrhu historických databází, analýzy a kódování dat.[117] Spekulovalo se rovněž o založení studijního oboru historická výpočetní věda, k čemuž sice nedošlo, nicméně v následku na to byla později (v roce 1985) založena asociace pod názvem Association for History and Computing pod vedením Petera Denleye a Deiana Hopkina. V jejím rámci se

[114] GREENSTEIN, D. Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States, s. 357.
[115] Tamtéž, s. 358.
[116] Tamtéž, s. 355.
[117] Tamtéž, s. 357.

v následujících dvou desetiletích každoročně konaly konference a vycházel časopis s názvem History and Computing.[118]

Tou dobou existovaly dva základní pohledy na to, jak by měla aplikace počítačů v historii vypadat. První z nich spočíval v přesvědčení, že počítačové metody jsou samy o sobě bez jakýchkoli dalších úprav dost dobré na to, aby byly v historii přímo aplikovatelné, přičemž druhý tvrdil opak a jeho zastánci byli toho názoru, že tyto metody musí být vždy specifickým způsobem upravené a vylepšené, aby mohly odpovídat požadavkům historického bádání.[119] Později se ukázalo, že první přístup, který vycházel z jednotící síle počítače, jenž je na základě metod schopen spojit různé vědní obory se v praxi neosvědčil kvůli jisté naivitě v jeho záměru.[120]

V 80. letech se objevovaly i názory, že historici by měli být víc explicitní při formulování svých vědeckých přístupů a strategií, aby byly tyto snadněji přeložitelné do počítačového jazyka. Jak bylo tou dobou všeobecným trendem, i zde bylo možno nalézt přesvědčení, že tato schopnost by měla být součástí výbavy každého historika a že každý z nich by měl být schopný používat programovací jazyk pro tvorbu programů.[121] Pochopitelně bylo možné narazit i na opačné tendence v podobě varování před tím, aby historické bádání nebylo podřízeno počítačové metodě, která by měl být pouze jednoduchým nástrojem. Jinými slovy, aby kvůli „pohodlnosti“ pro počítač nebyly projekty upravované pro jeho potřeby.[122]

Od konce 70. let se pro práci s historickými projekty začaly používat tzv. relační databáze. Tento přístup k informacím byl rozšířený mezi britskými historiky více než mezi historiky kontinentálními.[123] Jedná se o soubor vzájemně propojených tabulek obsahujících určitá data, které slouží jako základní způsob reprezentace pramenů. Data uložená v databázích mohou být různé povahy: čísla, text, zvukový záznam nebo obraz. Zpravidla jsou data vysoce strukturovaná a umístěna do série menších numerických a textových položek (tzv. pole) spojených do tabulek.[124]

[118] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 27.
[119] Tamtéž, s. 28-29.
[120] Tamtéž, s. 30.
[121] Tamtéž.
[122] Tamtéž, s. 30-31.
[123] Tamtéž, s. 31-32.
[124] HARVEY, Ch., PRESS, J. Databases in Historical Research, s. 10.

Přesto se první pokus o speciální úpravu technologie relačních databází pro historické účely objevil v Německu. V roce 1980 Manfred Thaller, zaměstnanec Max Planck Instut für Geschichte v Göttingenu, vyvinul speciální systém pro správu historických databází – CLIO. Ačkoli byl tento systém primárně určený pro výzkum probíhající v samotném institutu, byl zpřístupněný i pro použití v jakýchkoli jiných historických projektech.[125]

Thallerův systém CLIO nebyl příliš úspěšný, především pak kvůli tomu, že nebyl dostatečně uživatelsky přívětivý. Z jeho pohledu je totiž využívání komputačních metod v historii komplikovaná záležitost a kdyby to chtěli historici zjednodušovat, mohli by sami sebe připravovat o užitečné nástroje. Tento přístup se ukázal jako nepříliš vhodný a ve výsledku tento vývoj vedl k tomu, že v průběhu 90. let začaly vznikat další, přístupnější systémy.[126]

Jedním z nich byl software vyvinutý norským historikem Janem Oldervollem pro analýzu historické demografie na základě dat ze sčítání lidu v Norsku během devatenáctého století. Tento systém se jmenoval CensSys a umožňoval vyhledávání a některé další operace s daty (např. vytvoření skupin podle místa bydliště nebo věku) v databázi, která počítala zhruba tři miliony záznamů.[127] V Nizozemsku poté Leen Breure vyvinul systém Socrates, což byl nástroj pro vytváření historických databází. Obsahoval i návody pro historiky, jak s ním mají pracovat a návrhy pro strukturování databází.[128]

Mezi další velké databáze patří například soubor obchodních záznamů a právních dokumentů britské Východoindické společnosti, která umožňuje vyhledávání podle typu záznamu. Většina dokumentů je uložena v podobě fotokopie, nicméně pro některé z nich je umožněno textové vyhledávání, buď v plné nebo omezené míře.[129] Dále pak například kompletní korespondence Viléma Oranžského (přibližně 13 000 dopisů), která bohužel není kompletně digitalizovaná – zpřístupněná je v podobě krátkých anotací s metadaty obsahujících odkazy na fotokopie dopisů a dalších písemností, které jsou v analogové podobě uloženy ve dvou set různých archivů a knihovnách po celé Evropě.[130] Také

[125] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 27.
[126] Tamtéž, s. 34.
[127] OLDERVOLL, J. A System For Analysing Census-Type Data, s. 17, 21.
[128] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 35.
[129] Dostupné z: http://www.eastindiacompany.amdigital.co.uk/Introduction/Guide
[130] Dostupné z: http://resources.huygens.knaw.nl/wvo/en

americká Knihovna Kongresu provozuje databázi korespondence George Washingtona, která počítá kolem šedesáti tisíc záznamů v podobě fotokopií, některé z nich jsou opatřeny i přepisem.[131]

Dodnes se uživatelé používající knihovní nebo archivní katalogy setkávají s technologií relačních databází, byť v uživatelském rozhraní často nejsou spoje patrné. Na rozdíl od minulosti, kdy se k tomuto účelu používalo buď zmiňované CLIO nebo Microsoft Access, dnes většinou tyto databáze fungují pomocí dotazovacího jazyka SQL (nejpoužívanější variací je MySQL).[132]

Dalším směrem vývoje digitálních metod v evropské historiografii byla práce s obrazy, které byla v roce 1985 dokonce věnovaná první samostatná konference. Tou dobou se totiž jednak vyvíjely geografické informační systémy (o nichž se blíže zmíníme dále), ačkoli později tyto metody přezvali spíše archeologové než historici, a jednak probíhala digitalizace obrazů jako takových, např. rakouská akademie věd už na konci 70. let zahájila projekt, který byl věnován digitalizaci tisíce různých obrazů zachycujících každodennost a materiální kulturu pozdního středověku, přičemž obrazy byly po digitalizaci opatřeny metadaty, mimo jiné i textovými popisy pro jednodušší vyhledávání. Od začátku 90. let počet podobných projektů neustále narůstá.[133]

Samostatným vývojem, značně nezávislým na historii, procházelo používání počítačových metod v archeologii. Poměrně dlouhou dobu po svém vzniku se totiž počítače v rámci této disciplíny skoro vůbec nepoužívaly, jenom výjimečně pro statistické zpracování, což pochopitelně nedělá základ archeologické práce. Počítače jako nástroje pro uchovávání archeologických záznamů začaly nabývat na významu teprve v polovině 70. let se vznikem relačních databází. Databáze umožnily uchovávání záznamů o archeologických nálezech v podobě, která byla pohodlnější pro zpracování než klasický ruční způsob. Archeologové byli schopni rozpoznat užitečnost počítačů pro svou práci poměrně rychle; v roce 1973 se v Anglii na University of Birmingham začali scházet odborníci se zájmem o tuto problematiku, což položilo základ asociaci s názvem

[131] Dostupné z: https://www.loc.gov/collections/george-washington-papers/about-this-collection/
[132] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 31-32.
[133] GREENSTEIN, D. Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States, s. 358.

Computer Applications and Quantitative Methods in Archaeology. V roce 1984 také vznikl časopis Archaeological Computing Newsletter.[134]

Mimo jiné spousta odborníků v oblasti archeologie ocenila možnost snadno nahlédnout do záznamů jiných pracovišť bez nutnosti absolvovat jakoukoli cestu. Spojení záznamů z různých projektů ale zůstává i v současné době spíše cílem než reálnou skutečností, protože dodnes v oboru neexistují všeobecně uznávané standardy uložení dat. Situace se navíc komplikuje i tím, že se výrazně liší potřeby samotných archeologů a pracovníků muzeí, což mnohdy vede k nekompatibilnosti databází vytvořených na různých pracovištích. I přesto ale existují některé souhrnné databáze věnované archeologickým vykopávkám v jisté oblasti, např. *The Archaeological Settlements of Turkey* nebo *National Archaeological Database of US public archaeological sites* nebo projekt *OASIS* v Británii.[135]

Existuje ale i kritika používání počítačových metod v archeologii, která se zakládá především na tvrzení, že archeolog, který se spoléhá na digitální databáze, riskuje, že ztratí vazbu na reálné nálezy, protože s nimi nebude ve fyzickém kontaktu.[136]

Digital history v období webových projektů:

Předchůdcem moderních digitálních projektů je *The Philadelphia Social History Project* započatý již v roce 1969 (ukončen 1985) na University of Pennsylvania pod vedením Theodora Hershberga. Cílem projektu bylo hlubší porozumění průběhu urbanizace a industrializace na materiálu Philadelphie mezi lety 1850-1880. Tou dobou bylo město druhé největší v Americe a působilo jako jedno z největších průmyslových center. Pro účely výzkumu byla vytvořena databáze vedená na magnetických páscích uchovávající záznamy o dvou a půl milionech lidí, kteří žili ve městě ve druhé polovině 19. století a rovněž informace o ubytování, podnicích, továrnách a dopravních prostředcích. Projekt rovněž pracoval s geografickými daty – každý obyvatel byl přiřazen k určité čtvrti ve městě. Práce byla zaměřena na několik podoblastí – povaha pracovních aktivit v továrnách, použití městského prostoru, základní průběh života obyvatel (události jako svatba, stěhování, změna práce atp.) a postavení etnických menšin se zvláštním zřetelem na irskou a německou komunitu. Archiv vytvořený pro tento projekt nebyl

[134] EITELJORG, H. Computing for Archaeologists, s. 21-22.
[135] Tamtéž, s. 22-23.
[136] Tamtéž, s. 23.

zpřístupněn veřejnosti, badatelé jej používali pro svůj vlastní výzkum, načež publikovali pouze výstupy těchto výzkumů. Projekt je důležitý zejména vzhledem k faktu, že se v jeho době jednalo o největší digitální archiv vůbec.[137]

Právě otázka zpřístupnění historického výzkumu širší veřejnosti se ukázala jako klíčová pro rozvoj moderní digitální historie: přelomovým bodem se totiž, jak jsme již zmínili, stal vznik internetu. Prvním internetovým historickým projektem (a ostatně i jednou z prvním webových stránek vůbec) je *The Valley of the Shadow Project* spuštěný v roce 1991 na University of Virginia, jehož autory jsou americký historici William G. Thomas III z University of Nebraska–Lincoln a Edward Lynn Ayers, současně působící na University of Richmond. Tento projekt odstartoval působení Institute for Advanced Technology in the Humanities (IATH).[138]

Jedná se o webové stránky přibližující zkušenosti vojáků konfederace z oblastí Augusta County a Viriginie a vojáků unie z Franklin County a Pensylvánie. V samotném úvodu projektu stojí, že „v tomto digitálním archivu můžete prozkoumat tisíce originálních dopisů a deníků, novin a projevů, záznamů sčítání lidu a církevní záznamy". Lze zde nalézt zmapovaná tři období: „The Eve of War", trvající od podzimu 1859 do jara 1861, „The War Years" mezi lety 1861 a 1865 a následně „The Aftermath" od jara 1865 do podzimu 1870. Mapa, která tato tři období znázorňuje, úmyslně připomíná místnosti domu či knihovny, aby byla přístupnější pro uživatele a obsahuje odkazy jako galerie obrazů, databáze dopisů a deníků, databáze tisku, vojenské záznamy, daňové záznamy apod. V databázích je nadále umožněno vyhledávání pomocí různých parametrů, např. fotografie je možné vyhledávat podle témat, jmen zachycených osob nebo míst.[139]

Webové zpracování nebylo původním záměrem, ale v roce 1993 se tvůrci seznámili s možnostmi internetu pomocí prvního webového prohlížeče Mosaic a okamžitě se rozhodli pro výraznou změnu celého projektu. Do konce roku už byla hotová první pracovní verze webových stránek, ačkoli zpočátku nízká rychlost internetového připojení bránila potenciálním uživatelům v seznámení se s daty v tom rozsahu, v jakém

[137] HERSHBERG, T. The Philadelphia Social History Project: An Introduction, s.43-44.
[138] Dostupné z: http://valley.lib.virginia.edu/VoS/usingvalley/valleystory.html
[139] Tamtéž.

to tvůrci projektu zamýšleli – kupříkladu nebylo možné načítat obrázky. Proto byl projekt v roce 1994 vydán ještě ve formátu CD-ROMu.[140]

Stránky se postupně zdokonalovaly, zároveň se zvyšovala i rychlost internetového připojení, která dovolovala uživatelům nahlédnout všechna zpřístupněná data. Finální podoba stránek byla dokončena v roce 2007, čímž byl projekt uzavřen. Celkově bylo zpřístupněno kolem 12 000 souborů opatřených metadaty.[141]

Projekt byli především zpočátku schopni plně ocenit zejména učitelé historie jako skvělý nástroj pro popularizaci, zatímco samotní historici si zachovávali spíše skeptický odstup a nevnímali tento digitální archiv jako něco mimořádného. Postupně se ale mínění historické komunity obrátilo a projekt začal být vyzdvihován především kvůli tomu, že byl schopen vykreslit velice komplexní obraz občanské války, který by nebylo možné postihnout v plné šíři za použití tradičních médií.[142]

Ve stejné době vznikl v City University of New York další velký historický projekt pod vedením Roye Rosenzweiga s názvem *Who built America?*, digitální databáze obsahující videozáznamy, audionahrávky, fotografie a mapy. Projekt byl vydán v podobě CD-ROMu, což, jak se později ukázalo, byla velká nevýhoda, protože na rozdíl od *The Valley of the Shadow* se tento projekt nikdy nestal známým široké veřejnosti, zatímco počet návštěvníků The Valley dosáhl několika milionů.[143]

Dalším významným projektem je *The Proceedings of the Old Bailey – London's Central Criminal Court 1674 to 1913*. Jedná se o sbírku všech dochovaných soudních záznamů z trestního soudu Old Bailey (centrální část Londýna), přičemž celkový počet zpracovaných řízení dosahuje téměř 200 000. Texty jednání jsou plně přepsané a digitalizované a opatřené metadaty. Na základě tohoto je uživatelům umožněno vyhledávání v celé databázi podle zadaných parametrů jako jsou například datum, jména, pohlaví a sociální postavení účastníků, jména soudců a samozřejmě kategorie zločinu. Pro novější texty byl vyvinut nástroj pro automatickou analýzu, starší jsou zpracovány ručně. Projekt je provozován na The Digital Humanities Institute, který přináleží The University of Sheffield.[144]

[140] Tamtéž.
[141] Tamtéž.
[142] THOMAS, W.G. Computing and the Historical Imagination, s. 62.
[143] Tamtéž, s. 62-63.
[144] Dostupné z: https://www.oldbaileyonline.org/static/Project.jsp

Na University College of London dlouhodobě funguje pracoviště The Centre for Editing Lives and Letters (CELL), které se zabývá digitalizací především raně novověké korespondence. Mezi jinými zde můžeme nalézt dopisy Francise Bacona, Roberta Boylea nebo Thomase Bodleye. Texty jsou zpřístupňovány v přepsané podobě, ale jsou vždy také opatřeny fotokopií originálního pramene.[145] V současné době se pracoviště věnuje například výzkumu barokní latiny či projektu ve spolupráci s John Hopkins University *The Archaeology of Reading in Early Modern Europe*, který zkoumá dynamiku vývoje čtení a čtenářů v raně novověké Evropě – v rámci tohoto projektu bude zpracován korpus 36 plně digitalizovaných raně novověkých tisků opatřených desítkami tisíc ručně psaných poznámek z pera Johna Dee a Gabriela Harveyho.[146]

Dalším projektem probíhajícím na Roy Rosenzweig Centre je například *Making the History of 1989 – The Fall of Communism in Europe*. Jedná se o digitální sbírku pramenů různé povahy týkajících se pádu komunismu ve východní Evropě, mezi něž patří literární texty (například Havlova *Moc bezmocných*), dobové fotografie, karikatury, stranické dokumenty či dopisy; nechybí ani statistiky, týkající se například porodnosti nebo základních makroekonomických indikátorů. Každý pramen je opatřen poměrně rozsáhlým popisem a obsahuje seznam klíčových slov pro snadnější vyhledávání. Všechny prameny jsou přeloženy do angličtiny, ačkoli ne vždy jsou takto přístupné v celé podobě (některé je možné nahlédnout v podobě abstraktu nebo stručného popisu).[147]

V oblasti digitalizace a 3D modelace je významným počinem projekt *Rome Reborn*, který, byť je primárně popularizační zaměřený na širokou veřejnost a výuku historie, přesně zobrazuje v 3D modelu, jak vypadal starověký Řím co do architektury i městského plánování. Podrobně jsou zde vykresleny nejznámější památky jako Koloseum, Forum Romanum, Pantheon a další. Projekt běží již od poloviny 90. let a v současné době se pracuje na třetí verzi vizualizace, která se připravuje pro virtuální realitu.[148]

[145] Dostupné z: http://www.livesandletters.ac.uk/projects
[146] Dostupné z: http://www.livesandletters.ac.uk/projects/archaeology-reading-early-modern-europe
[147] Dostupné z: http://chnm.gmu.edu/1989/
[148] Dostupné z: https://www.romereborn.org/content/aboutcontact

Metody a nástroje (a jak je správně používat)

Co se týče způsobů digitální reprezentace historických pramenů, existují tři základní možnosti: jedna z nich je digitální faksimile, což je naskenovaná kopie psaného nebo tištěného pramene s přepisem a informací o osobách, místech a občas i konceptech zmíněných textů; druhá forma je digitální edice, která na rozdíl od faksimile zpřístupňuje několik různých verzí stejného pramene, jako příklad může sloužit edice všech psaných a tištěných vydání Chaucerových Canterburských povídek, což dělá dohromady 84 rukopisů a 4 vydání před začátkem 16. století – jinými slovy, v prvním případě je to vždy jedna a ta samá verze konkrétního pramene, kdežto v druhém se jedná o soubor různých verzí stejného pramene, který umožňuje zkoumat jeho proměny např. v čase. Třetí možností je pak celý digitální archiv nebo digitální knihovna, jako jsou *Archivo General de Indias*, což je archiv státních dokumentů španělské koloniální administrace v Americe nebo *Duderstadt Project*, který obsahuje digitalizovanou verzi městského archivu, či *Fontes Civitatis Ratisponensis*, dokumenty týkající se středověkého Řezna. Archivy se ovšem nemusejí týkat jen textů – příkladem může být *Prometheus*, archiv obrazů z různých muzeí, určených především pro potřeby dějin umění.[149]

Ačkoli pro historiky je nejvhodnějším způsobem zobrazení pramenů edice, je jasné, že většina historických textů nikdy nebude v této podobě zpracována. I když se knihovny a archivy snaží digitalizovat co největší počet zdrojů, značný rozsah práce zůstává na historicích samotných. Z toho plynou omezení toho, co a jak bude zpracováno vzhledem k časovým a finančním hlediskům.[150]

Při vytváření historických databází či nástrojů musíme mít neustále na paměti několik základních skutečností. Jedna z nich je, že aby měla jakákoli data smysl, musí být doplněna interpretací, takže je potřeba tyto interpretace zapracovat do databáze samotné nebo do dalšího digitálního nástroje. Zároveň ale interpretace nesmí ovlivňovat samotnou podobu tohoto nástroje. Jinými slovy, tým, který připravuje projekt, musí především myslet na to, aby nevnucoval svou vlastní interpretaci dat databázi samotné – ta by měla být oproštěna od jakýchkoli hypotéz, aby případní další badatelé, kteří by snad s touto databází či nástrojem v budoucnu dále pracovali, mohli aplikovat své vlastní domněnky

[149] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 41-43.
[150] Tamtéž, s. 45-46.

a interpretace bez toho, aby byli samotnou povahou nástroje nuceni je upravovat podle jeho předchozího nastavení.[151]

Je také potřeba rozlišovat mezi historickou informací a syrovými historickými daty. Aby se z dat stala informace, musí být data vybraná, editovaná, popsaná, reorganizovaná a zpřístupněná v podobě, která umožňuje historické bádání.[152]

Takto strukturována data jsou zpravidla uchovávána v databázích. Neustále probíhají diskuze o tom, jestli by měla historická databáze nějakým způsobem odpovídat struktuře pramenů, které zaznamenává, a pokud ano, jak lze vyřešit problémy s tímto spojené. Některé druhy pramenů je totiž těžké reprezentovat pomocí relačních databází, protože různé prameny stejného typu můžou mít odlišnou datovou strukturu, a dokonce občas i v rámci jednoho pramene nemusí být datová struktura jednotná – v rámci např. jedné kroniky se může lišit struktura údajů co do zaznamenávání data, měn, vlastních jmen, popř. může být i jedna kronika psaná odlišným pravopisem či dokonce jazykem.[153]

Peter Denley poukazuje na několik základních metodologických problémů, s nimiž se může historik potýkat při vytváření databáze:

1. Komplexní struktura pramene, která je často obtížně převoditelná do počítačové podoby a musí tedy v některých případech být obětována kvůli snadnějšímu zpracování
2. Hierarchická struktura pramene, která znesnadňuje vytváření návrhu databáze
3. Nejednotnost textových záznamů, která nutí v některých případech k uchovávání prázdných dat, což zpomaluje počítač (tento problém byl aktuální v dřívějších dobách, kdy výpočetní kapacita byla mnohem menší, v dnešní době ho sotva můžeme považovat za relevantní)
4. Neúplnost historických dat
5. Nepřesnost historických dat
6. Obtížná převoditelnost formátů (např. různé standardy pro záznam času, měr nebo měn).[154]

[151] Tamtéž, s. 20.
[152] Tamtéž, s. 21.
[153] Tamtéž, s. 47.
[154] DENLEY, P. Models, Sources and Users: Historical Database Design in the 1990s, s. 34-35.

Všechny tyto faktory nemálo znesnadňují tvorbu systematicky strukturované databáze. Právě toto napětí mezi komplexitou historických dat a pevnou strukturou počítačových databází bylo jedním z hlavních témat diskuzí probíhajících v 90. letech. Nutnost výběru mezi velkým množstvím pramenů a mezi omezenými časovými a finančními zdroji způsobilo vznik dvou opačných názorů ke zpracování historických databází – přístupu z pramene a přístupu z modelu.[155]

Přístup, který vychází z modelu, především strukturu databáze zakládá na tom, že má dopředu definovaný systém tabulek, který umožní snadné vyhledávání a až následně se snaží pramen převést do podoby oněch tabulek. V případě, kdy vycházíme z pramene, je přístup opačný – napřed se zachycuje původní text v plném rozsahu, ideálně s potřebným kontextem a struktura databáze se vymýšlí až později zpětně.[156]

Nejznámějším zastáncem druhého přístupu je Manfred Thaller (zmiňovaný v předchozí části jako tvůrce CLIO), který na konci 80. let zformuloval několik základních doporučení pro vytváření historických databází. Vzhledem k tomu, že historie jako taková se potýká s jistými specifickými problémy, které nejsou přítomné v jiných disciplínách – jako neúplnost či nepřesnost pramenů, neurčitost a nemožnost jednoznačné identifikace autora – musí na rozdíl od klasických projektů stát u zrodu takové databáze vždy historik, který může nabídnout odborné znalosti. Dále tvrdí, že reprezentace dat by měla vycházet z podoby a povahy daného pramene, nikoli z architektury programu. Třetí doporučení pak vychází z přesvědčení, že systém by měl spojovat databázi s plnotextovým přístupem k pramenům, což znamená, že bude obsahovat kompletní text daného pramene a zpřístupní tento text uživateli, ideálně v kombinaci s určitou metainformací, např. v podobě historického kontextu. Historický nástroj také musí počítat s nejednotností pravopisu a umožňovat přenos dat do formátu, který bude umožňovat statistickou analýzu.[157]

Tento přístup založený na přesvědčení, že jakýkoli nástroj by měl především vycházet z podoby reprezentovaného pramene, se ovšem setkal s rozsáhlou kritikou. Bylo

[155] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 47, 58.
[156] THOMAS, W.G. Computing and the Historical Imagination, s. 60.
[157] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 33-34.

mu především vytýkané možné přeceňovaní pramene a absence interpretací, které jsou pro racionální kritiku pramenů klíčové.[158]

Mezi nejznámější zastánce opačného přístupu patří Lou Burnard a Charles Harvey, kteří tvrdí, že při vytváření databáze by měli historici nejdříve navrhnout konceptuální model, který odráží objekty a události v reálném světě spolu s jejich vzájemnými vztahy a až následně by měl být model zmapován pomocí počítačových datových struktur.[159]

Denly navíc poukazuje na to, že přístup vycházející z vytváření modelů nikoli z pramenů je vhodnější pro kvantitativní analýzu většího počtu běžných pramenů. Druhý přístup je potom naopak vhodnější, pokud se zpracovává malé množství pramenů specifické povahy. Na výběr metodologie má podle autora vliv i prostředí, z něhož historik pochází: zatímco britští historici tradičně inklinují v modelově-orientovanému přístupu, němečtí častěji upřednostňují databáze založené na struktuře pramene.[160]

Další problém spočívá v tom, že rozdíl mezi těmito dvěma přístupy je velký a rozhodnutí musí být učiněno v raných fázích projektu. Pochopitelně by byl vhodnější postup, který by umožnil základní zpracování dat bez toho, aby se těmto datům musela hned ze začátku přiřadit pevná struktura. Takový postup ale v současné době běžně není technologicky zvládnutelný.[161]

Požadavky, kterým by ideálně měla odpovídat historická databáze, jsou takové: měla by vystihovat komplexitu původního pramene, být dělitelná na různé moduly čili drobnější části, které se dají zpracovat samostatně, odpovídat určitým vybraným standardům, umožňovat výběr vhodných metod a nebýt příliš časově náročná na vypracování.[162]

Dobrým příkladem zpracování historické databáze je knihovna *Perseus*, která obsahuje dokumenty v řečtině, latině, angličtině, italštině a arabštině. Jedná se většinou o literární texty z různých období, přičemž zpracování těchto textů postupovalo od

[158] DENLEY, P. Models, Sources and Users: Historical Database Design in the 1990s, s. 39.
[159] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 32-33.
[160] DENLEY, P. Models, Sources and Users: Historical Database Design in the 1990s, s. 40.
[161] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 49.
[162] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 50.

jednoduchého lingvistického značkování k vypracování komplikovanějších metadat. Příprava databáze probíhala poloautomatickým způsobem ve dvou krocích: prvním byla automatická analýza a detekce vlastních jmen, druhým potom následná manuální úprava dat na základě toho, co předzpracoval počítač. V rámci projektu vznikl automatický systém analýzy slovních tvarů v řečtině, latině a italštině, byly vytvořeny nástroje pro vyhledávání v textech, pro propojení textů odkazy, pro získání místních názvů z textů a následné vytváření map, vyhledávání dat jako časových údajů a následné promítnutí těchto údajů na vytvořené časové osy.[163]

Z počítačové lingvistiky se také do historie dostala metoda zkoumání tematické koncentrace, o níž budeme podrobněji pojednávat v druhé části práce. Největší problém při aplikaci této metody v historii je práce s tématy, která přetrvávají napříč dlouhými časovými úseky. Tato metoda byla totiž původně vytvořena především pro analýzu velkého počtu menších textů, především novinových článků z omezeného časového úseku, zatímco v historii je často nutné pracovat s menším počtem delších textů, které jsou od sebe vzdálené v čase.[164]

Co se týče statistických metod, jejich používání v historii se v průběhu času měnilo, v 70. a 80. letech v období rozmachu kliometrie byla statistika používaná především k testování hypotéz, stejným způsobem jako v sociologii nebo ekonomii. V současné době se statistika v historii častěji chápe jako popisný nástroj čili spíše prostředek k tomu, abychom byli schopni rozpoznat určité vzorce ve velkém množství historických pramenů.[165]

Jednou z těchto statistických metod je vícenásobná regrese, která zkoumá variaci jedné proměnné, např. růstu počtu obyvatel konkrétního místa v konkrétní době v závislosti na několika nezávislých faktorech jakými mohou být stav obchodu v okolí, míra rozvoje zemědělských technologií, počet válečných střetů a dalších. Tato metoda je nicméně v oblasti bádání v historii poměrně málo používaná, v podstatě se využívá jen takto v oblasti historické demografie nebo dějinách politických systémů (vývoj voličských preferencí atp.).[166]

[163] Dostupné z: http://www.perseus.tufts.edu/hopper/research
[164] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 52.
[165] Tamtéž, s. 58.
[166] Tamtéž, s. 59.

Jako schůdnější cesta pro statistický výzkum se může ukázat regrese založená na menším vzorku a menším počtu proměnných. Příkladem může sloužit výzkum prováděný na univerzitě v Lundu, kdy se zkoumal dopad hospodářské krize na porodnost ve Švédsku v 19. století na základě dat získaných ze čtyř farností.[167]

Dalším přístupem je tzv. historie událostí či analýza historických událostí (event history analysis) založená na zkoumání toho, s jakou pravděpodobností se v životě konkrétního jedince odehrála nebo mohla odehrát jistá událost, jako např. vstup do manželství, změna bydliště, porod dítěte, úmrtí v mladém věku, získání dědictví… K tomuto účelu slouží speciální software neboli TDA (transitional data analysis software) vytvořený již v roce 1997 ve Spojených Státech.[168]

Ke zkoumání vývoje jisté proměnné v závislosti na čase slouží tzv. analýza časových řad. Možné aplikace můžeme najít v rámci historické demografie, např. ve výzkumu počtu nemanželských dětí v 19. století v Anglii (podle prováděného výzkumu klesal), při sledování vývoje volební účasti či v oblasti historické klimatologie. Dále je možné pochopitelně aplikovat tuto metodu pro potřeby ekonomické historie, jako např. v případě výzkumu prováděném v roce 1998 na sledování cen obilí v Anglii mezi lety 1450-1812.[169]

Mezi další vhodné metody patří tzv. data mining. Od předchozích se liší především schopností odhalit skryté vlastnosti datového souboru, na jejichž základě je teprve poté sestavena testovatelná hypotéza. Nejpoužívanějším nástrojem data miningu je hierarchické shlukování čili automatické rozdělení datasetu na skupiny s podobnými měřitelnými vlastnostmi. Výsledkem může být buď shlukový graf nebo tzv. stromový graf, dendrogram. Často se shluková analýza používá v kombinaci s geografickým informačním systémem (GIS; viz dále).[170] Tímto přístupem se budeme více zabývat v kapitole věnované strojovému učení.

Na používání statistiky je založená i metoda počítačové simulace, která se snaží předpovídat vývoj konkrétního systému podle zadaných vstupních parametrů, pro jejichž vývoj se vytváří statistický model. Jednodušší simulace, která nebere v potaz předchozí stavy systému, se označuje jako tzv. Markovův řetězec, komplikovanější, založená na

[167] Tamtéž, s. 60.
[168] Tamtéž.
[169] Tamtéž, s. 61-62.
[170] Tamtéž, s. 63-65.

teorii pravděpodobnosti, se poté nazývá metoda Monte Carlo. Příkladem použití této metody může sloužit poloautomatická simulace propuknutí první světové války nebo simulace ustálení hranic evropských států, která poukázala na vliv geografických faktorů na dřívější ustálení hranic v západní Evropě oproti východní. Simulace se používá především v hospodářské historii a historické demografii, mimo tyto disciplíny je použití spíše ojedinělé.[171]

Co se týče práce s obrazy, ještě před nedávnem bylo jejich používání pro historický výzkum velice omezené. Zatímco v popularizačních historických pracích vždy hrál obraz velkou roli, v samotném bádání se mu nepřikládal velký význam – většinou byl vnímán jako pouhá ilustrace. Dnes se ovšem situace značně mění, protože technologie umožňují snadné zpracování a šíření obrazu; budeme-li se odkazovat na Petera Burka, zvyšuje se zájem o kulturu obrazů a obraz se začíná vnímat jako historický pramen. Obraz se stává součástí informačního systému historie stejně jako textová nebo číselná data.[172]

S grafickými a multimediálními daty pochopitelně souvisí metoda vizualizace, což je názorná reprezentace statistických zjištění. Od klasické reprezentace se přitom liší tím, že nejenže tato zjištění zpřístupňuje, ale zároveň umožňuje nový náhled na data. Stejně jako data mining nám i vizualizace může pomoci najít souvislosti, které vzhledem k počtu dat dosud zůstávaly skryty.[173]

Další metodou související s grafikou je GIS neboli geografický informační systém; jedná se o spojení historie jakožto disciplíny studující časovou diferenciaci s geografií jakožto disciplínou studující prostorovou diferenciaci. Jinými slovy, jde se o počítačový systém, který propojuje geografické souřadnice s časovou osou. GIS může pomoci odpovídat na otázky typu, zda se zkoumaný předmět někam nepřemístil nebo nezměnil svoje hranice v určeném časovém úseku. Používání tohoto systému je běžnější v archeologii než v historii – v dnešní době je zde GIS spíše pravidlo než výjimka.[174]

Mimo vytváření databází pracují archeologové rovněž s digitálním modelováním terénu a vytvářením map. Zatímco v 70. letech se jednalo o ojedinělé případy, od poloviny 80. let se spolu se vznikem geografického informačního systému začaly tyto modely a mapy vytvářet stále častěji. Princip GIS zde spočívá v propojení klasické relační

[171] Tamtéž, s. 66.
[172] Tamtéž, s. 69-70.
[173] Tamtéž, s. 73.
[174] Tamtéž, s. 14, 95.

databáze, kde může být například uložená informace o různých nálezech, flóře nebo fauně, s body na mapě a s geografickou informací, např. strmost terénu v určité oblasti. GIS tedy umí propojit jednotlivé body na mapě s informacemi uloženými v databázi.[175]

Dalším grafickým nástrojem používaným v zejména archeologii je tzv. počítačem podporované projektování (computer aided design, CAD), které se používá od poloviny 80. let. Používá se k modelování, v současné době i trojrozměrnému, míst archeologických vykopávek a nálezů. Lze jej uplatnit dvojím způsobem: buď pro modelování konkrétních objektů v nepozměněné podobě anebo pro digitální rekonstrukce poškozených nebo neúplných projektů; v některých případech se může jednat i o celé městské panorama (viz například již zmiňovaný projekt *Rome Reborn*). Největším nedostatkem této metody je, nebo tomu tak alespoň do nedávné minulosti bylo, nedostatečná výpočetní kapacita k vykreslení veškerého potřebného kontextu pro znázornění zkoumaného objektu. Dalším problémem poté je, že obecně programy používané pro účely archeologie jsou původně vyvíjené ke komerčním účelům a nejsou tedy vždy hodné a přizpůsobené potřebám archeologického bádání.[176]

Současný stav, problémy a perspektivy:

„*Historik, který považuje počítač za neužitečný, tím ignoruje velkou část historického bádání a nadále nebude brán vážně*." (Onno Boonstra, 1990).

Tento výrok byl (jistě v nadsázce) napsán s nadějí, že s rozvojem počítačů nejenže se objeví nové oblasti historického bádání, ale že počítače budou schopny nabídnout řešení mnohých problémů, kterými se historie jako věda vyznačuje. Realita je ovšem taková, že tyto problémy přetrvávají dodnes. Jsou to především potíže textového charakteru, jako co přesně znamená konkrétní slovo, proč tam je, proč byl napsán konkrétní text, kdo byl jeho autor, kdo měl být čtenář či jak se stalo, že se zrovna tento text dochoval. Dále existují problémy s identifikací, například jak si můžeme být jistí, že osoba, o níž referuje jeden z pramenů, je tatáž osoba, o které se zmiňuje druhý? Následují problémy se strukturací dat – jakým způsobem budeme ukládat v databázích metadata například o historickém kontextu či problémy vizualizace: jak promítneme na mapu

[175] EITELJORG, H. Computing for Archaeologists, s. 23-24.
[176] Tamtéž, s. 24-26.

parametry, které se mění v čase, jako např. počet obyvatel konkrétních míst nebo jejich národnostní složení.[177]

I když se nedá říct, že se aplikace počítačových metod v historii vůbec nevyvíjí (naopak, v posledních letech vznikají stovky různých projektů), nemůžeme pořád tento rozvoj označit za probíhající uspokojivým tempem. Boonstra říká, že vývoj je pomalý nikoli kvůli tomu, že by počítače neměly vhodné nástroje, jež by mohly nabídnout historii, nýbrž proto, že samotná historie nebyla schopná tyto nástroje v dostatečné míře akceptovat.[178]

Velké množství historických prací již teď můžeme označit prakticky za vzniklé pomocí počítačů, protože jsou založeny na výzkumu zdrojů, které byly napřed převedeny do podoby databází.[179]

Historická informační věda se ale netýká jenom bádání, ale rovněž i zpřístupnění historických pramenů. Čím dál více se objevují další a další digitální archivy, archivní zdroje se převádí do digitální podoby, navíc edice pramenů, které byly v minulosti publikované v tištěné podobě, se čím dál více objevují v podobě elektronické. Sice je v současné době jen zlomek kulturního dědictví uloženého v různých institucí přístupný v elektronické podobě, nicméně tento počet neustále roste.[180]

V dnešní době se již nezpochybňuje, že elektronické zdroje jsou velice důležité pro zpřístupnění pramenů, které by jinak zůstaly ukryté, nepřístupné a nedostupné k analýze. Digitální média jsou nepochybně vhodnější pro zveřejnění zdrojů než papír, v mnoha ohledech také než mikrofilm. Není proto divu, že spousta vydavatelů a institucí se obrací čím dál častěji k digitálním zdrojům.[181]

Přesto přístup k datům nadále zůstává jedním z klíčových problémů humanitních věd. Prvním krokem k vyřešení tohoto problému je zmíněná digitalizace katalogů knihoven, archivů a muzeí a jiných institucí zpřístupňujících kulturní dědictví. Pořád ale existuje v podstatě nekonečné množství potenciálně relevantních pramenů pro humanitní výzkum, které se současně nachází v analogové (nedigitální) formě a mohou být digitalizovány. Vzhledem k tomuto neomezenému množství je nevyhnutelné, že

[177] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 9.
[178] Tamtéž.
[179] Tamtéž, s. 14.
[180] Tamtéž, s. 19.
[181] Tamtéž, s. 13.

digitalizace bude selektivní, částečná – vždy se digitalizuje jen konkrétní část celku, přičemž zbytek zůstane nedotčen.[182]

Další příležitosti historického výzkumu se otevírají spolu se zpřístupněním velkého množství elektronických obrazů v podobě fotografií nebo maleb. Jak už bylo řečeno, historici tradičně používají obraz pouze jako ilustrační materiál, ale takto zpřístupněné velké kolekce obrazů mohou být podrobeny přímé analýze. Nehledě na to pořád narůstá počet pramenů, vznikající rovnou v digitální podobě, jako např. e-maily, které se v budoucnu nepochybně také stanou předmětem historického bádání.[183]

Překážkou pro rozvoj digitálních metod je i skutečnost, že komunita historiků, která se všeobecně nevyznačuje přílišnou láskou k počítačům, byla dlouhodobě od komputačních metod odrazovaná i uživatelskou nepřívětivostí a dlouhou dobou, již vyžaduje učení se těmto metodám. Na druhou stranu zde údajně hrálo roli i nepochopení a nedostatek zájmu z jejich strany.[184]

Aby historie mohla odpovídat novým požadavkům digitální éry, měli by historici být více informačně gramotní, ale rovněž by měli i informatici lépe chápat specifika historie. V souvislosti s tím se občas tvrdí, že nová generace historiků by měla mít jak znalosti v oblasti historie, tak i pokročilé informační dovednosti jako programování nebo vývoj databází.[185]

V oblasti historie pořád přetrvává rozdíl mezi menšinou historiků, kteří používají počítače jako nástroje pro analýzu historických dat a převládající většinou, jež je používá pouze k běžným kancelářským účelům. Dalším problémem je, že relevantní výzkumy týkající se aplikace počítačových metod v historii jsou často prováděny odborníky, kteří sami nejsou historici. Toto způsobuje používané metody nejsou vždy dost dobře přizpůsobeny pro historické účely a zároveň, že historici nejsou dostatečně informováni o dostupných metodách nebo nástrojích.[186]

Problematická je ovšem i počítačová dovednost archeologů, protože schopnosti, které se používají při archeologické práci, především schopnost navrhnout databázi a pracovat s CAD nebo GIS, nejsou v dostatečné míře podporované na univerzitách.

[182] Tamtéž, s. 14.
[183] Tamtéž, s. 19.
[184] Tamtéž, s. 50.
[185] Tamtéž, s. 14.
[186] Tamtéž, s. 86.

Nadále také přetrvává absence všeobecně uznávaných standardů práce s daty v oboru – problém, který můžeme spatřovat napříč celou sférou digital humanities.[187]

V současné době existuje paradox, kdy pochopitelně počítač sám od sebe není schopen automaticky nabídnout všechny potřebné nástroje, a tudíž pro vznik nových metodologických přístupů a nového náhledu na prameny je potřeba, aby byly tyto nástroje napřed vyvinuty. Problémem zůstává, jak jsme již zmínili, že samotní historici jsou málokdy schopni je vytvořit, zatímco odborníci, kteří toho schopní jsou, nemají dostatečnou představu o potřebách historie. Z toho ovšem neplyne, že každý historik by měl být počítačovým expertem; nicméně pokud takových historiků nebude dostatečný počet, disciplína zůstane odkázaná na použití komerčních programových řešení se všemi jejich omezeními. Na druhou stranu, co se týče nedostatečného zájmu informatiků o historické metody, je tato skutečnost poměrně nepřekvapivá, jelikož i samotná historická komunita je rozdělená a nemůže se shodnout na tom, zda nějaké pokročilé počítačové metody vůbec potřebuje. Pokud ale historici budou pokračovat v odmítání moderních technologií, hrozí, že dojde k jistému odtržení od celkového vývoje humanitních věd.[188]

Co se týče vyhlídek do budoucna, musí digitální historie vykonat jisté nezbytné úkoly – například vyvinout systém získání, zpracování, ukládání a značkování dat, upravit statistické metody pro použití v historii nebo přizpůsobit multimediální nástroje pro zveřejnění výsledků výzkumu.[189]

Rosenzweig směrem k budoucnosti disciplíny říká, že „problémy, kterým se historici budou věnovat, budou v zásadě stejné jako problémy současné historiografie, zatímco metodologie se pravděpodobně výrazně změní.[190] Budoucnost digitální historie by měla ležet především ve vytváření nástrojů a technik, které nejsou běžně součástí „tradiční“ historické práce, zejména s přihlédnutím k onu faktu, že zdroje, s nimiž historici pracují, se budou do budoucna stále více objevovat v elektronické podobě, ať už se jedná o e-maily, videa, digitální fotografie či textové soubory. Dále by se zájem měl přesunout z prostého nahlížení na pramen na celkový proces, „dělání“ digitální historie.[191]

[187] EITELJORG, H. Computing for Archaeologists, s. 28.
[188] BOONSTRA, O., BREURE, L., DOORN, P. Past, Present and Future of Histocial Information Science, s. 91-92.
[189] Tamtéž, s. 94.
[190] THOMAS, W.G. Computing and the Historical Imagination, s. 64.
[191] SEEFELDT, D., THOMAS, W.G. What is Digital History? In: Perspectives in History, 2009, č. 5.

Aplikace počítačových metod při práci s historickými prameny

V následující části práce se pokusíme blíže zaměřit na některé z metod, jež tradičně spadají do oblasti počítačové a kvantitativní lingvistiky a poukázat na to, jakým způsobem je lze využívat, pracujeme-li s nimi jako s nástrojem pro analýzu historického pramene. Kriticky důležitou součástí takovéhoto přístupu je ovšem dosud ne vždy snadno překonatelný problém – všechny texty, se kterými pracujeme, musí být ve formátu přístupném počítači, což u mnoha pramenů v současné době není a nemůže být standardem. Proto pro naše potřeby předvádíme postupy na souboru pramenů, jež jsou v již přepsané podobě dostupné z webových stránek Letters of William Herle Project. Tento projekt vznikl v rámci pracoviště Centre for Editing Lives and Letters (CELL), o němž jsme se již zmiňovali v předchozí části práce.[192]

O Williamovi Herlem samotném toho nevíme příliš mnoho, není například známo datum jeho narození. Patrně měl základy právnického vzdělání, jazykový talent (údajně mluvil latinsky, francouzsky, italsky, německy a holandsky) a jisté společenské konexe v nejvyšších kruzích anglické elity (je například znám jako chráněnec Roberta Dudleyho, hraběte z Leicesteru), což ho predisponovalo k diplomatické službě, kterou vykonával pro královnu Alžbětu I. až do své smrti v roce 1588, přičemž většinu své kariéry strávil v Nizozemsku, odkud pravidelně informoval svého patrona hraběte z Leicesteru, královskou radu nebo i samotnou královnu o událostech a náladách v souvislosti s nizozemským povstáním.[193]

Postupně se tedy v následujících kapitolách na příkladu Herleho dopisů pokusíme přiblížit, jak může historik postupovat, pokud se rozhodne využívat digitální metody a jaký přínos, stejně jako možné nevýhody, může takovýto postup mít.

Metody kvantitativní lingvistiky

V obecném popisu toho, co je vlastně kvantitativní lingvistika jako celek, se můžeme vymezit oproti lingvistice klasické, jež zkoumá především obecnější pravidla jazyka. Kvantitativní lingvistika se naopak zajímá spíše o konkrétní texty a jejich vlastnosti, přičemž je důležité si uvědomit, že to, co nazýváme vlastnostmi, není inherentní, tj. jedná se vždy o naše konstrukce. Vlastnosti textu jsou z tohoto pohledu

[192] *Letters of William Herle Project*. Dostupné z: http://www.livesandletters.ac.uk
[193] Dostupné z: http://www.livesandletters.ac.uk/herle/introduction.html

měřitelné pomocí konkrétních přístupů, kvantitativní lingvistika se na jejich základě snaží formulovat obecné hypotézy, které později testuje na dalších textech.[194]

Z pohledu historie ovšem jsou důležité pouze samotné výsledky měření, které nikdy nemohou stát samy – vždy je kriticky nutné podložit měření kvalitativní interpretací dat, bez níž by tyto údaje ztrácely smysl. Veškeré metody kvantitativní lingvistiky tedy mohou sloužit pouze jako pomocný nástroj při standardním historickém bádání.

Web scraping a vytváření databáze

Základem pro jakékoli zkoumání je jako v ostatních případech mít připravený soubor textů ve formátu .txt, v našem případě soubor Herleho osobní korespondence – jednak dopisy, jejichž autorem je on sám, jednak ty, jejichž je recipientem. Vytvořit soubor lze buď manuálně jednoduchým kopírováním textu z webových stránek, nebo použít metodu tzv. web scrapingu, která automaticky uloží zvolený obsah stránek do počítače uživatele. Podobné skripty bývají nejčastěji založeny na knihovně Scrapy v programovacím jazyce Python.

Pro účely naší práce byl použit tento:

```
import scrapy
import re
import requests
from bs4 import BeautifulSoup

def get_page(link):
    page = requests.get(link)
    page = page.content
    soup = BeautifulSoup(page, 'lxml')
    text = str(soup.find('div', {'class': 'letter'}))
    text = re.sub(r'<.*?>', '', text)
    text = re.sub(r'&amp;', '&', text)
    text = re.sub(r'\r\n', ' ', text)
    text = re.sub(r'\n', '', text)
    text = re.sub(r'\[.*?\]', '', text)
    text = re.sub(r'\s\s+', ' ', text)
    text = re.sub(r'^', '', text)
    return text

class HerleSpider(scrapy.Spider):
    name = 'herle'
    allowed_domains = ['www.livesandletters.ac.uk']
```

[194] ČECH, R., POPESCU, I., ALTMANN, G. *Metody kvantitativní analýzy (nejen) básnických textů*, s. 8-12.

```
    start_urls = ['http://www.livesandletters.ac.uk/herle/recipient_author.html']

    def parse(self, response):
        recipients = response.xpath("//h2//text()").getall()
        lists = response.xpath("//ul").getall()
        id = 0
        for i in range(len(lists)):
            letters = lists[i].split('<li class="item">')
            for letter in letters:
                if len(letter) > 14:
                    id += 1
                    recipient = recipients[i]
                    author = re.search(r'(?<=<strong>).*(?=</strong>)',
letter).group(0)
                    year = re.search(r'15.*?(?=\.)', letter).group(0)
                    year = year.replace('[', '')
                    year = year.replace(']', '')
                    year = int(re.search(r'\d\d\d\d', year).group(0))
                    url = re.search(r'letters.*\.html', letter).group(0)
                    link = response.urljoin(url)
                    text = get_page(link)

                    yield {
                        "id": id,
                        "url": link,
                        "author": author,
                        "recipient": recipient,
                        "year": year,
                        "text": text
                        }
```

Pomocí takto definovaného kódu lze nahradit ruční vyhledávání a ukládání textů. Skript postupně projde seznam dopisů zveřejněný na stránkách a stáhne je do počítače uživatele. Následně pro každý z jednotlivých dopisů v tomto seznamu stáhne a uloží jeho obsah, přičemž budou všechny umístěny v jednom .json souboru (jenž je v současné době prakticky standard pro ukládání internetových dat), v němž bude pro každý dopis uloženo jeho identifikační číslo, v našem případě podle pořadí, ve kterém se dopisy stahovaly ze stránek, následně odkaz na dopis, jméno odesílatele, jméno adresáta, rok a samotný text dopisu. Obsah uložený ve formátu .json lze následně otevřít a nahlédnout například v programu notepad++.

Následně byl za pomoci dalšího skriptu obsah tohoto .json souboru uložen do databáze SQlite, která byla pro tyto účely vytvořena. Výhodou tohoto typu relačních

databází je, že nepotřebují pro své fungování žádný internetový server, jsou jednoduché na spravování a jsou uloženy v podobě jednoho souboru na lokálním počítači uživatele. Tento soubor se pak dá otevřít i pomocí jakéhokoli vhodného manageru databází (programu, který umí databáze zobrazit), v našem případě byl použit DB Browser (SQlite).[195]

Tato databáze byla vytvořena pomocí následujícího skriptu v Pythonu za použití dotazovacího jazyka SQL (na skriptu v trojitých uvozovkách):

```
import sqlite3
import json

conn = sqlite3.connect('herle.db')
db = conn.cursor()
db.execute(
    """ CREATE TABLE IF NOT EXISTS letters (
                                        id integer PRIMARY KEY,
                                        url text NOT NULL,
                                        author text NOT NULL,
                                        recipient text NOT NULL,
                                        year integer,
                                        letter text
                                    ); """
)

with open ("herle.json") as my_file:
    data = my_file.read()
herle_dict = json.loads(data)

for i in range(len(herle_dict)):
    row = herle_dict[i]
    insert = (row["id"], row["url"], row["author"], row["recipient"], row["year"],
row["text"])
    db.execute(
    """ INSERT INTO letters
        VALUES (?,?,?,?,?,?); """, insert)
conn.commit()
db.close()
```

Jedná se o jednoduchou databázi, která se skládá jen z jedné tabulky, přičemž sloupce v tabulce odpovídají datům vytaženým z internetu (oněch šest zmiňovaných

[195] Naše databáze dopisů je k nahlédnutí uložena na adrese: https://drive.google.com/drive/folders/1xiMTHjHu9OclmNayuiClQlTMp5kZnk1d?usp=sharing

vlastností) a řádky představují každý jednotlivý dopis. Výhodou takto vytvořeného databázového zobrazení je, že máme všechny dopisy zpřístupněné na jednom místě, takže není nutné každý z 303 dopisů ukládat jako samostatný textový soubor. Zároveň DB Browser umožňuje filtrování podle obsahu jednotlivých polí, což znamená, že máme možnost zobrazit například všechny dopisy z konkrétního roku, všechny dopisy od konkrétního odesilatele nebo s konkrétním příjemcem, popřípadě kombinace těchto filtrů (například všechny dopisy Herleho královně Alžbětě z roku 1583) atd.

Frekvence, frekvenční slovník a tematická koncentrace

Klíčovým pojmem kvantitativní lingvistiky je pojem frekvence. Obecně se pod tímto termínem chápe počet opakování jisté události v určitém časovém úseku. Z lingvistického pohledu se pak jedná o množství výskytu nebo četnost určitého jevu.[196] Altmann frekvenci definuje jako „počet výskytu slova v určitém úseku textu" a rovněž poukazuje na to, že „frekvence slov je jednoduchá vlastnost, ležící takříkajíc na povrchu textu a je tak k dispozici nejenom lingvistům, ale i nelingvistům". Výsledky získané měřením se pak mohou používat v různých oborech včetně typografie, psychologie, psychiatrie, výuky jazyků, kryptografie či vývoje softwarových aplikací.[197] Jak ovšem pracovat s frekvencí z pohledu historie?

Pracujeme-li s korpusem různorodých textů, které mohou být různých žánrů či odlišné délky, je třeba používat namísto absolutní frekvence tzv. relativní. Relativní frekvence znamená, že neřešíme reálný výskyt slov, ale přepočítáváme onen výskyt vzhledem k velikosti textu, v němž se slovo nachází.[198] Relativně schůdným řešením je používat místo absolutní frekvence procentuální zastoupení, které vyjadřuje, jaká část textu je tvořena opakováním zkoumaného slova. Příkladem použití může být situace, kdy zkoumáme dva prameny odlišné délky, přičemž v obou pramenech potřebujeme zkoumat výskyt stejného slova. V kratším z obou textů bude z pochopitelných důvodů absolutní frekvence slova nižší – procentuální zastoupení ovšem vyjádří jeho reálnou váhu se zohledněním délky textu.

Základním nástrojem, pomocí kterého lze pracovat s frekvencemi, je tzv. frekvenční slovník. Jedná se o seznam slov seřazených sestupně podle jejich frekvence

[196] CVRČEK, V. *Frekvence*. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy – Nový encyklopedický slovník češtiny. 2017. Dostupné z: https://www.czechency.org/slovnik/FREKVENCE.
[197] POPESCU, I. *Word Frequency Studies*, 2009, s. 14.
[198] CVRČEK, V. *Frekvence*.

v textu. Seznam zpravidla mívá podobu tabulky, v jejímž prvním sloupci je tzv. *rank* čili pořadí podle četnosti výrazu. Poté následuje samotná frekvence vyjádřená přesným číselným údajem (kolik daných slov se reálně v textu vyskytuje), případně procento relativní frekvence. Ve třetím sloupci je pak uvedeno samotné slovo.

Jistou komplikaci způsobuje skutečnost toho, že základní podoba frekvenčního slovníku nezohledňuje fakt, že stejné slovo se může vyskytovat v různých slovních tvarech. To může být problém zejména při práci s jazyky mající výraznou flexi (skloňování substantiv a adjektiv a časování sloves) jako jsou čeština, němčina nebo latina, což jsou právě jazyky, se kterými historik v českém prostředí pracuje nejvíc. K vyřešení tohoto problému klasicky slouží tzv. lemmatizéry, tj. jsou programy nebo skripty, které automaticky zařadí všechny slovní tvary do základní podoby, a ve frekvenčním slovníku se poté neuvádí všechny tvary daného slova, ale pouze tato tzv. lemmata. Lemmatizéry existují v současné době na množství jazyků, a to včetně klasických či některých dalších mrtvých jazyků.[199]

Ačkoli jsou lemmatizéry často volně přístupné na internetu, jejichž použití vyžaduje mnohdy pokročilé uživatelské znalosti v oblasti počítačové lingvistiky, přičemž se badatel mnohdy dostane do situace, kdy i existující lemmatizéry nefungují dost přesně na právě ten jazyk pramenů, se kterým v danou chvíli pracuje. Vždy je nutné počítat s možností, že určitý počet slovních tvarů lemmatizér nemusí vždy rozpoznat.

K vytvoření samotného slovníku slouží různé nástroje, z nichž pro uživatele nejpřístupnější jsou již hotové lingvistické aplikace jako je například AntConc, či program Quita vyvinutý na Katedře obecné lingvistiky Univerzity Palackého v Olomouci. Další je možností je vytvoření vlastního skriptu za pomoci jednoho z programovacích jazyků, přičemž výhoda tohoto přístupu spočívá ve větší flexibilitě a možnosti program upravit podle potřeb konkrétního textu. Nejpoužívanější jazyk je Python, kde pro účely kvantitativní analýzy textu existuje rozsáhlá knihovna NLTK.

Jako příklad slouží následující skript napsaný právě v jazyce Python:

```
def freq (text):
    from decimal import Decimal

    def tokenize_text (text):
```

[199] ČECH, R., GARABIK, R., ALTMANN, G. Testing the Thematic Concentration of Text. In: *Journal of Quantitative Linguistics*, č. 3, 2015, s. 218.

```
        import re
        return re.split("\W+", text.lower())

    def remove_blank_tokens (tokens_list):
        return list(filter(lambda x: x != "", tokens_list))

    tokens_list = remove_blank_tokens(tokenize_text(text))
    output = ""

    freq_list = {}
    for token in tokens_list:
        if token not in freq_list:
            freq_list[token] = 1
        else:
            freq_list[token] += 1

    rank = 0

    for item in sorted(freq_list, key = freq_list.get, reverse = True):
        rank += 1
        rel_freq = round(Decimal(freq_list[item]/len(tokens_list))*100, 3)
        output += str(rank) + "\t" + str(freq_list[item]) + "\t" + str(rel_freq) +
"\t"  + item + "\n"
    return output

with open ("input.txt") as my_file:
    text = my_file.read()

with open ("output-freq.txt", "w") as txtfile:
        txtfile.write(freq(text))
```

Tento skript na základě regulárních výrazů rozdělí text uložený v souboru input.txt na jednotlivé části, tzv. tokeny. Tokeny se rozumí jakékoli arbitrárně určené jednotky, v našem případě konkrétní slova. Následně je spojí do seznamu, z něhož pak udělá frekvenční slovník (s absolutní i relativní frekvencí v procentech). Ten uloží v podobě textového dokumentu output-freq.txt do stejné složky, v níž se nachází skript.

Výslednou tabulku přikládáme k práci v příloze č. 1.

Jak si lze povšimnout, na prvních místech tabulky se vyskytují tzv. synsémantika neboli stopwords. Jedná se o neplnovýznamová slova jako jsou spojky, předložky, spony apod. Plnovýznamová slova jako substantiva či verba, jež se nazývají autosémantika, jsou v tabulce umístěna dle ranku níže. Jakousi pomyslnou hranici mezi oblastí synsémantik a autosémantik tvoří tzv. h-bod, jež se určuje následujícím způsobem:

Zjistíme, kde v tabulce se nachází slovo, jehož rank se rovná jeho frekvenci. Vždy záleží na rozsahu textu, u větších textů bude h-bod z přirozených důvodů níže. Pokud se vyskytne situace, kde není možné najít přesnou shodu mezi rankem a frekvencí, za h-bod se považuje hranice mezi dvěma nejvhodnějšími slovy. H-bod je důležitým indikátorem, jestli má daný text výrazné tematické zaměření.[200] Slova, která se nachází nad touto hranicí, výrazně odráží zaměření textu, a jsou proto nazývána jako slova tematická.[201]

V našem frekvenčním slovníku vytvořeném z dopisů (v příloze zkráceném po h-bod) je prvním tematickým slovem „good" na 37 místě. Následují „majestie" s rankem 40 „sayd" s rankem 53, „master" s rankem 63, „monsieur" s rankem 80, „made" s rankem 90, „men" s rankem 97, „matter" s rankem 99, „herle" s rankem 105, „england" s rankem 106, „god" s rankem 107, „state" s rankem 109, „frenche" s rankem 112, „place" s rankem 149, „honor" s rankem 151, „hand" s rankem 156, „favor" s rankem 157, „prince" s rankem 159, „towne" s rankem 160 či „end" s rankem 177.

Z takto vyčísleného seznamu můžeme získat jistý přehled např. o tom, kdo může být častým adresátem dopisů či o čem může Herle psát. Byť mnohé ze slov mohou mít pochopitelně zdvořilostní charakter (např. oslovení – „majestie", „master", „monsieur", či fráze - „honor", „god" apod.), můžeme odvozovat alespoň některá z témat, kterým se dopisy věnují, např. podle slov „england" lze hádat, že Herle sleduje anglické státní zájmy, dle „frenche" může zase pozorovat aktivity nebo kroky spojené s Francií nebo že „towne" se budou týkat situace holandských měst. „State" a „matter" zase naznačují, že se probírají státní záležitosti, „favor" naopak může odkazovat k záležitostem osobním.

Zajímavá skutečnost nicméně je, že například slovo „frenche" se nad h-bod dostalo, nicméně například jiná slova odkazující obecně na nizozemské prostředí už ne, jako např. „low countries", „holland", „orange" či množství místních názvů. Na základě toho můžeme vytvořit hypotézu, že Herle se více zajímal o mezinárodní politiku (a především logicky zájmy Anglie), než o vnitřní záležitosti v Nizozemsku. K tomu, abychom však mohli tuto hypotézu potvrdit či vyvrátit, už by ale bylo nutné přejít ke kvalitativnímu výzkumu založeném na klasických historických metodách.

[200] POPESCU, I. Text Ranking by the Weight of Highly Frequent Words. In: Exact Methods in the Study of Language and Text (ed. GZYBEK, P., KÖHLER, R.). Berlin, 2007, s. 559.
[201] ČECH, R., POPESCU, I., ALTMANN, G. *Metody kvantitativní analýzy (nejen) básnických textů*, s. 15-17.

V souvislosti se zkoumáním pramene pomocí tematických koncentrací textů můžeme sledovat i proměny používání jednotlivých slov. Tento přístup nám může pomoci ve větším souboru textů identifikovat proměny používání konkrétního pojmu v závislosti na čase nebo prostoru. Vhodnou aplikací této metody je sledování politického či sociálního diskurzu, např. vývoj používání ideologické terminologie v totalitním tisku, nicméně vždy je záhodno využít metod korpusové lingvistiky.

V našem případě znázorníme proměny četnosti užívání některých tematických slov v korespondenci v čase. Pro tuto část je nutné rozdělit velký soubor všech textů na menší úseky, z nichž každý bude obsahovat dopisy napsané v konkrétní rok.

Tuto úlohu vykoná následující skript:

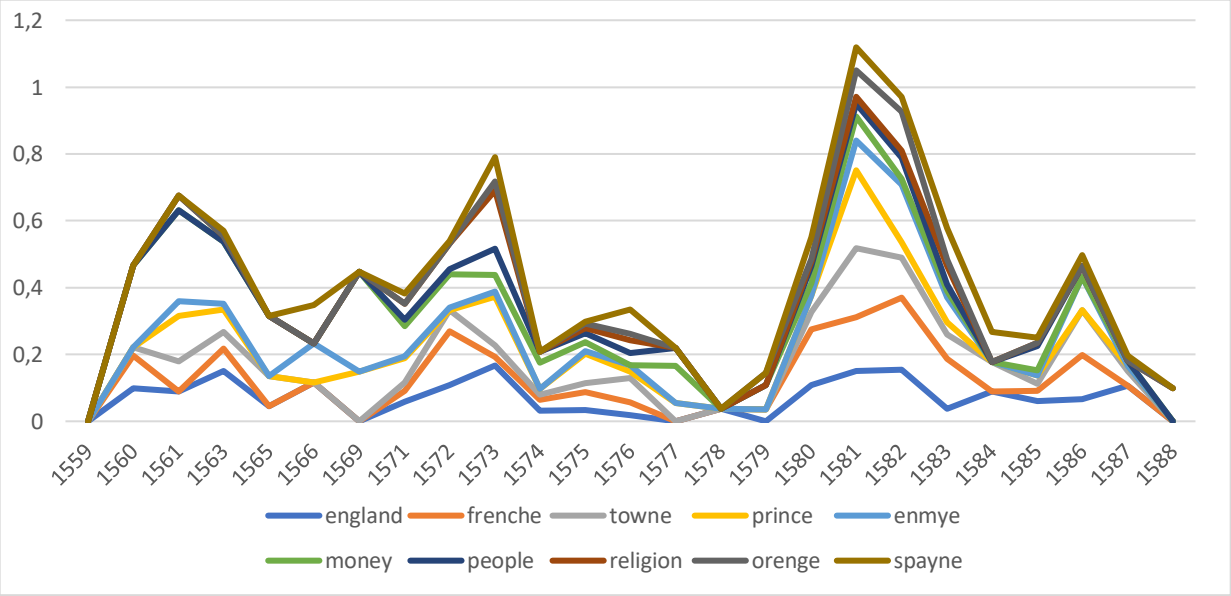
```
import json

with open ("FromToHelre.json") as my_file:
    data = my_file.read()
texts = json.loads(data)

years_list = []
for text in texts:
    if text["year"] not in years_list:
        years_list.append(text["year"])
years_list = sorted(years_list)

for year in years_list:
    year_text = ""
    for text in texts:
        if text["year"] == year:
            year_text += "\n" + text["letter"]
    with open ("{}.txt".format(year), "w") as txtfile:
        txtfile.write(year_text)
```

Následně pro každý z vytvořených souborů vygenerujeme vlastní frekvenční slovník, který bude využit k tomu, abychom mohli porovnávat používání jednotlivých slov v dopisech z různých let. K tomu slouží tento skript:

```
def CompareFrequencies (word):
    import glob

    output = word + "\n"

    file_list = [txt_file for txt_file in glob.glob("*.txt")]

    for my_file in file_list:
```

```
        with open(my_file) as input_file:
            freqlist = list(zip(*(line.strip().split('\t') for line in input_file)))

        for i in range(len(freqlist[0])):
                if freqlist[3][i] == word:
                    output += (my_file + '\t' + freqlist[1][i] + '\t' +
freqlist[2][i] + "\n")
                    break
        else:
            output += (my_file + '\t' + str(0) + '\t' + str(0) + "\n")

    return output
word = input("Type a word to compare its frequency: ")
print(CompareFrequencies(word))
```

Výsledkem je tabulka zachycující absolutní a relativní frekvence jednotlivého slova v konkrétních letech. Slova, jejichž proměny zaznamenáváme, nejsou slova čistě tematická (vyskytující se nad h-bodem), ale i zastoupení dalších pojmů, jejichž frekvence byla nižší. Jedná se o slova, jejichž užívání se mohlo změnit v závislosti na vývoji historických událostí. Příkladem může sloužit následující graf, který znázorňuje vztah mezi rokem a relativní frekvencí:



Jak je patrné, nejvyšší výskyt pojmů obecně sledujeme v letech 1581 a 1582. Tento fakt nepoukazuje ani tolik na reálnou skutečnost, jako znázorňuje častý problém, jemuž jsme při používání těchto metod vystaveni, a sice nerovnoměrnost množství vzorků, který nedokázalo zcela odstranit ani použití relativní frekvence. Především tedy u slov jako „spayne“, „orenge“, „religion“, „people“, „money“, „enmye“ a „prince“

nemůžeme s jistotou tvrdit, že jejich nárůst mezi lety 1580 a 1583 způsobilo jejich větší reálné užívání v Herleových dopisech, jako spíše musíme přijmout fakt, že data mohou být vinou nerovnoměrného množství textů pro daná léta zkreslená.

Je proto lepší nahlédnout na reálné množství dopisů, jež máme k dispozici. Naše databáze vypadá následovně:

Roky	Dopisy
1559	1
1560	1
1561	2
1563	8
1565	2
1566	1
1569	1
1571	29
1572	9
1573	9
1574	5
1575	17
1576	5
1577	2
1578	5
1579	3
1580	18
1581	22
1582	50
1583	8
1584	5
1585	27
1586	6
1587	16
1588	3

Srovnáním dat v tabulce a grafu nacházíme nespornou souvislost mezi velikostí databáze a výskytem slov v letech 1580, 1581 a především 1582. Na druhou stranu ovšem data z let 1571, 1575, 1585 a 1587 výrazné zvýšení užívání pojmů navzdory většímu množství dopisů nevykazují. Zaměříme-li se tedy na sledování výsledků s těmito skutečnostmi, můžeme formulovat následující hypotézy:

Existuje stabilita v užívání pojmů, jež je jeví jako klíčové (tematická slova), jako například „england“, jehož výskyt se bez ohledu na množství zachovaných dopisů v průběhu let výrazně nemění. Podobně, byť méně stabilní, je i výraz „frenche“. Lze se

tedy domnívat, že Herleova korespondence se ve sledovaném období vždy nějakým způsobem dotýkala francouzských a anglických zájmů. Naproti tomu spatřujeme výrazné výkyvy v užívání pojmů jako jsou „spayne“, „religion“ a „orenge“, které vykazují jistou míru korelace v letech 1573 či 1586. Byť vždy musíme mít na paměti, že korelace neimplikuje kauzalitu a že naše výsledky mohou být ovlivněny mnoha různými faktory včetně pouhé náhody, můžeme k případnému následujícímu kvalitativnímu výzkumu přistupovat s určitými předpoklady, jako například, že právě tyto roky zachycují jisté zajímavé historické skutečnosti, jako například že v roce 1573 Vilém Oranžský konvertoval ke kalvinismu či že byl vévoda z Alby odvolán jako nizozemský guvernér.

Je přirozené, že většina pramenů, s nimiž pracujeme, je zaměřena na určité téma či několik různých témat, přičemž v některých textech je toto zaměření výraznější než v jiných. Celkově se dá říct, že jak frekvenční slovník, tak tematická koncentrace mohou sloužit především k tomu, abychom byli schopni toto zaměření identifikovat dříve, než budeme pramen zkoumat podrobněji. Pochopitelně nemůžeme čekat, že by frekvenční slovník shrnul obsah textu jako takový (zejména zvážíme-li často hrozící nebezpečí, že nad h-bodem se nebudou vyskytovat autosémantika žádná – v tomto případě nám mohou sloužit jako jisté vodítko autosémantika těsně pod hranicí), ale lze s ním pracovat jako s pomůckou, která může sloužit jako první seznámení s jednotlivými texty či jejich souborem. Jinými slovy lze frekvenční slovníky využívat jako první krok při výběru relevantního pramene z většího množství textů.

Jako i u jiných metod kvantitativní lingvistiky je i zde kritickým bodem vhodný výběr pramene. Při práci s tematickou koncentrací je důležité využívat prameny, jež obsahují dostatečné množství textu, přičemž klíčové je zde užití substantiv. Takto zvolená metoda vyžaduje souvislý text, ideálně celý soubor textů, které tvoří větší celky, a to zejména v případě, je-li cílem dané texty mezi sebou porovnávat a zkoumat jejich proměny v čase atd.

Frekvenční slovníky lze používat i pro prameny ve formách seznamu, obsahující množství čísel a málo slov, které naopak nejsou příliš vhodné pro zkoumání tematické koncentrace. Frekvenční slovník nám tady může nicméně poskytnout cenné informace jako počet opakování jednotlivých položek např. při používání pramenů typu matrik, obchodních záznamů, soupisů majetku apod.

Korpusové metody

Další metodou původně náležící do oblasti lingvistiky je využívání tzv. korpusů. V širším smyslu se pod slovem korpus chápe jakýkoli soubor textů, přičemž jazykový korpus je poté specifická úprava tohoto souboru vytvořená za pomoci různých komputačních nástrojů, která v něm umožňuje rozšířené vyhledávání. V korpusu jsou tradičně jednotlivé texty nebo jednotlivá slova opatřena specifickou metainformací, jenž je uložena v podobě tzv. značky neboli tagu – text je v korpusu rozdělen na určité množství hierarchicky uspořádaných prvků, k nimž se tyto značky vážou. Můžeme například v rámci značek uchovávat informaci o autorovi či autorech textu, roku napsání nebo vydání; na nižších úrovních pak slovo zejména v lingvistických korpusech obsahuje jistou gramatickou informaci, např. slovní druh, pád, číslo, čas atd.[202]

Pro historické účely nenese pochopitelně takto uchovávaná gramatická informace příliš velký význam, proto je pro historické korpusy vhodnější vyvíjet specifické systémy značkování tradičně se zaměřením na vlastní jména nebo místní názvy; bylo by ale možné navrhnout i komplikovanější značkovací systémy, nabízí se využití šlechtických titulů, úřadů apod. – v tomto případě by, pokud by se takový způsob anotace osvědčil, existovala například možnost vyhledat všechny církevní tituly v daném korpusu. Takovýto úkol je nicméně nelehký a časově velmi náročný, proto pro potřeby této práce nebudeme vytvářet vlastním způsobem anotovaný korpus, namísto toho využijeme možností již zpracovaných programů, v našem případě Sketch Engine, jenž je přístupný na internetu.[203] Jedná se o grafické rozhraní pro korpusový manažer Manatee vyvíjený zaměstnancem Centra zpracování přirozeného jazyka Masarykovy univerzity Pavlem Rychlým ve spolupráci s britským lingvistou Adamem Kilgariffem [204].

Rozdíly mezi korpusem a databází

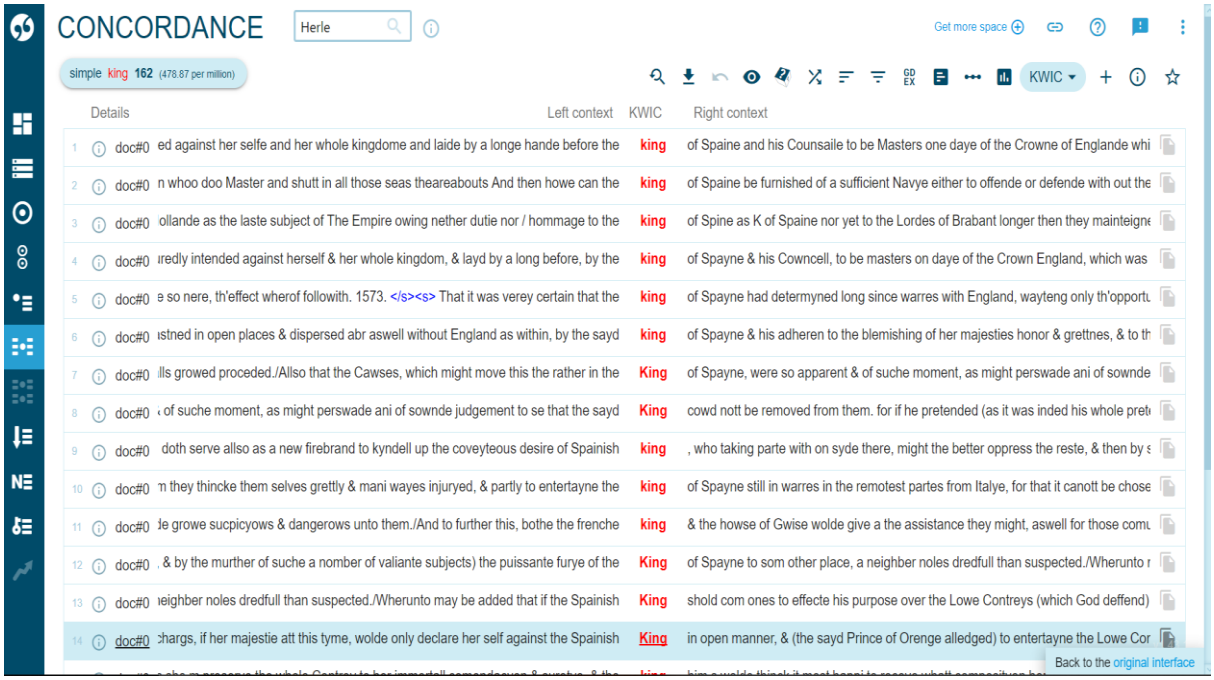
Korpus se od klasické databáze liší v první řadě svou strukturou. Databáze je tabulka či soubor vzájemně propojených tabulek, které obsahují pole různých tzv. datových typů (některé pole obsahuje text, jiné datum, jiné číslo atp.), zatímco korpus funguje jako uspořádaná hierarchická struktura jazykových jednotek různé velikosti, čímž rozumíme např. slovo, větu, odstavec, text a soubor textů. Veškerá metainformace je poté

[202] DASH, N. Corpus Linguistics: An Introduction, s. 27-31.
[203] Pro studenty UPOL a dalších českých akademických institucí zdarma do roku 2021.
[204] https://www.sketchengine.eu/sketch-engine-team/

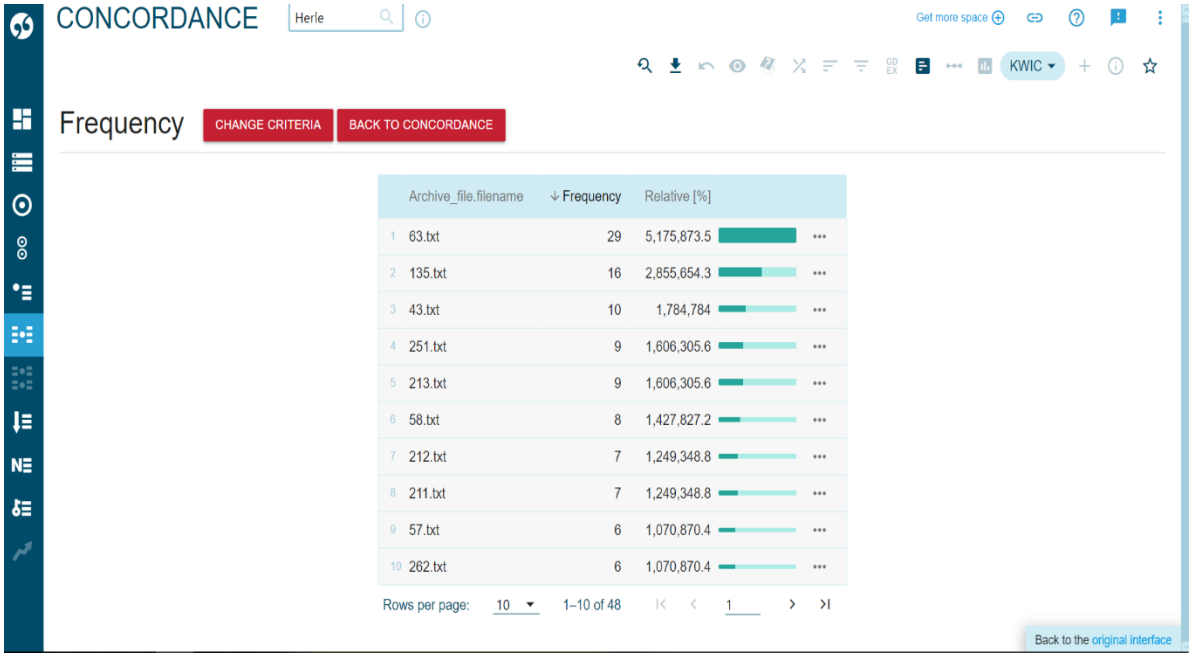
napojena na tyto prvky v podobě oněch značek. Ani účel, ke kterému jsou tyto nástroje využívány, není stejný – korpusy jsou používané primárně pro textovou analýzu, zatímco databáze slouží většinou k uchovávání záznamů různé povahy. Databáze nadále neumožňují vyhledávání slov pomocí klasických korpusových způsobů, jako je vyhledávání konkordancí, kolokací, n-gramů či klíčových slov.

Konkordance

Konkordance je, jak už jsme naznačili dříve, zobrazení hledaného slova v jeho bezprostředním kontextu, nejčastěji věty. V praxi s pomocí korpusového manažeru uživatel nejčastěji vyhledává zadané slovo a jako výsledek hledání pro něj systém připraví seznam všech konkordancí daného slova ve zkoumaném korpusu. V našem případě, pokud pracujeme s korpusem vytvořeným z Herleho dopisů, a vyhledáváme například slovo „king“, vypadá výsledek takto:

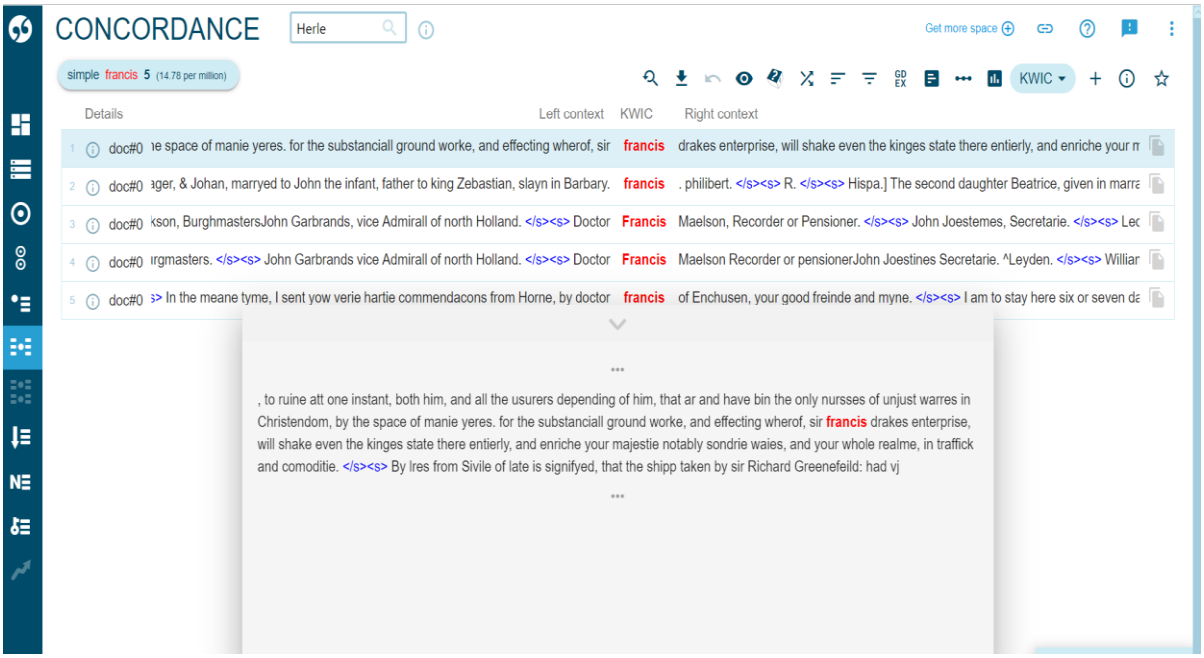


Jak vidíme, hledané slovo se objeví uprostřed stránky zvýrazněné červenou barvou. Z obou stran je slovo obklopeno kontextem (větou), ve kterém se vyskytuje. V případě potřeby širšího kontextu se dá slovo rozklepnout, aby se zobrazil celý odstavec. Vidíme, že výskytů tohoto slova je 162, přičemž relativní frekvence je 478 výskytů na milion. Vlevo je možné otevřít u každé konkordance přesnou lokaci čili soubor, ve kterém se tato věta nachází, a možnosti zobrazení frekvence tohoto slova napříč celým korpusem:



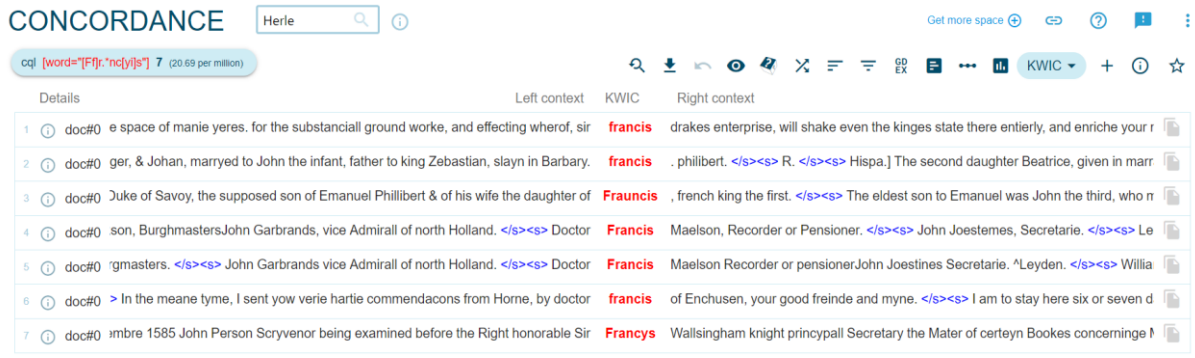
Zde vidíme deset textů, v nichž je hledané slovo „king“ obsaženo nejvíce z celkového počtu 48. Text č. 63, ve kterém je 29 výskytů, je dopis Williama Herleho adresovaný státníkovi Williamovi Burghleymu z roku 1573. Jedná se o dopis velmi obsáhlý, třetí největší z celého korpusu, což se nepochybně projevuje v absolutní frekvenci. Vysoká je ovšem i relativní frekvence, která nám může sdělit, že tento dopis je skutečně velmi výrazně zaměřený co do záležitostí týkajících se krále, v tomto případě dle konkordance zřejmě španělského.

V tomto bodě jsme tedy bez předchozí znalosti textů čistě za použití korpusu schopni odhalit výskyt určitého slova – pokud by nás tedy například zajímalo, zdali se v konkrétních textech vyskytuje jistá osoba, můžeme pouze zadat její jméno a podívat se, je-li toto jméno v textech obsaženo, nebo ne. Například, pokud by bychom hledali třeba jméno „Francis“, jednoduše jej vyhledáme tímto způsobem:



Zde vidíme všechny texty, ve kterých se dotyčné jméno vyskytuje, z čehož můžeme poznat, zdali se jedná o našeho hledaného Francise, nebo ne – pokud jsme tedy tímto mysleli například Františka z Anjou, můžeme takto shledat že ten, na rozdíl od třeba Francise Drakea, se v dopisech Williama Herleho nevyskytuje.

Tento způsob není ovšem stoprocentní, jelikož vzhledem k nejednotnému pravopisu, se kterým se obecně často v historických textech setkáváme (viz příloha č.1), je možné, že Francis nebyla jediná možná varianta tohoto jména, jakou lze v textech najít. Můžeme proto „pátrání“ rozšířit pomocí korpusového dotazovacího jazyka CQL[205].



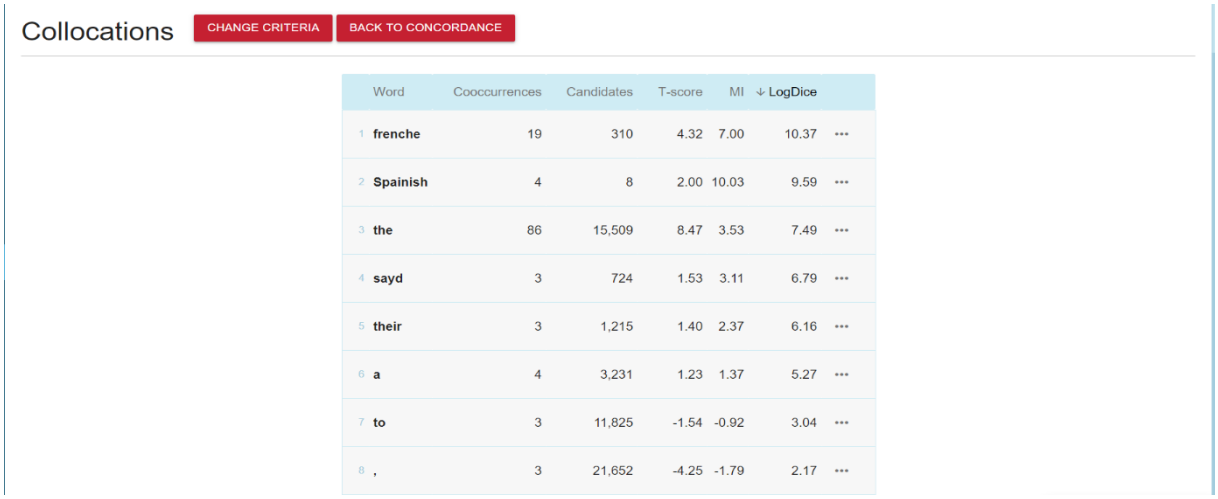
Zde můžeme pozorovat, že při zadání dotazu na různé „varianty“ jména Francis, které se liší co do pravopisu (vlevo nahoře) program vyhledá více výrazů než v předchozím kroku. Takovéto použití dotazovacího jazyka znamená, že hledáme slovo,

[205] Více o CQL na adrese https://www.sketchengine.eu/documentation/cql-basics/ nebo https://wiki.korpus.cz/doku.php/kurz:pokrocile_dotazy

jež začíná na velké či malé F (v hranatých závorkách), po R poté tečkou označujeme jakýkoli znak a hvězdičkou libovolný počet opakování přechozích znaků včetně nuly (to znamená, že se zde může vyskytnout „a“ stejně jako „au“) a následně pomocí dalších hranatých závorek mezi C a S vložíme, že hledáme buď měkké nebo tvrdé y. Tento dotaz je vhodné pokládat tak, aby byl s to pokrýt co největší množství různých pravopisných variant hledaného slova – v našem případě se tedy jednalo o Francis, Francys a Frauncis, přičemž pochopitelně existuje rozdíl mezi malým a velkým písmenem na začátku. Takto arbitrárně použitý pravopis nemusí (byť může) nutně poukazovat na různé autory či jiné skutečnosti, ale např. vzhledem k faktu, že různé varianty používá i sám Herle, často odkazuje spíše k neustálenosti psaného jazyka a neexistenci pravopisné normy v tomto období.[206]

Kolokace

Co se týká kolokací, jedná se o specifickým způsobem vyhledané slovo v textu, které se zobrazí spolu s dalšími slovy v jeho bezprostřední blízkosti. Od konkordancí se liší tím, že na rozdíl od kontextu ve větě při hledání kolokací najdeme jen ta slova, která

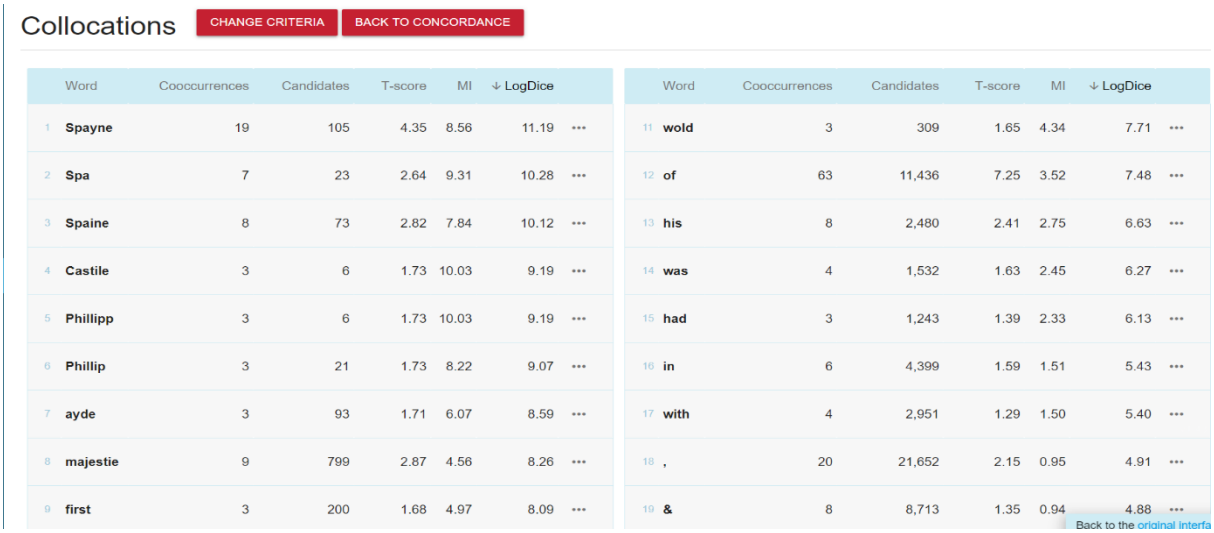


	Word	Cooccurrences	Candidates	T-score	MI	↓ LogDice
1	frenche	19	310	4.32	7.00	10.37
2	Spainish	4	8	2.00	10.03	9.59
3	the	86	15,509	8.47	3.53	7.49
4	sayd	3	724	1.53	3.11	6.79
5	their	3	1,215	1.40	2.37	6.16
6	a	4	3,231	1.23	1.37	5.27
7	to	3	11,825	-1.54	-0.92	3.04
8	,	3	21,652	-4.25	-1.79	2.17

se v textu pravidelně vyskytují v okolí slova zadaného. Jinými slovy, kolokace jsou slovní spojení, která se v textu pravidelně (více než jednou) opakují. Kolokace se vyhledává pro konkrétní slovo, přičemž vždy specifikujeme, zda hledáme pozici vpravo či vlevo od zadaného slova zároveň s vzdáleností od tohoto cílového slova. Pro historické účely je zde opět možnost využít tuto metodu při bližším seznamování se s velkým korpusem pramenů – pokud zadáme konkrétní slovo, v našem případě to může být například „king“,

[206] Burke, P. Languages and Communities in Early Modern Europe, s. 109.

program nám může pomoci zjistit, v jakém kontextu a o jakých králích se v pramenech nejčastěji mluví:



Collocations

CHANGE CRITERIA | BACK TO CONCORDANCE

	Word	Cooccurrences	Candidates	T-score	MI	↓ LogDice
1	Spayne	19	105	4.35	8.56	11.19
2	Spa	7	23	2.64	9.31	10.28
3	Spaine	8	73	2.82	7.84	10.12
4	Castile	3	6	1.73	10.03	9.19
5	Phillipp	3	6	1.73	10.03	9.19
6	Phillip	3	21	1.73	8.22	9.07
7	ayde	3	93	1.71	6.07	8.59
8	majestie	9	799	2.87	4.56	8.26
9	first	3	200	1.68	4.97	8.09

	Word	Cooccurrences	Candidates	T-score	MI	↓ LogDice
11	wold	3	309	1.65	4.34	7.71
12	of	63	11,436	7.25	3.52	7.48
13	his	8	2,480	2.41	2.75	6.63
14	was	4	1,532	1.63	2.45	6.27
15	had	3	1,243	1.39	2.33	6.13
16	in	6	4,399	1.59	1.51	5.43
17	with	4	2,951	1.29	1.50	5.40
18	,	20	21,652	2.15	0.95	4.91
19	&	8	8,713	1.35	0.94	4.88

Back to the original interfa

Zde vidíme dvojí vyhledání kolokací ke slovu „king“, přičemž první obrázek zobrazuje slova, jež stojí vedle tohoto výrazu přímo vlevo a druhý slova, která se objevují na druhé pozici vpravo (první pozice vpravo nebyla vyhledána z důvodu toho, že se na tomto místě zpravidla objevují pouze synsémantika, např. „of“) Z tohoto si můžeme udělat obrázek, že králové, kteří se v Herleho korespondenci objevují, jsou především králové Francie a Španělska. Je téměř jisté, že při takovémto zadání parametrů nehrozí šance, že by se „king“ častěji objevovalo spolu s jinými slovy, než jsou zde zobrazené – jedinou výjimkou může být zase různost pravopisu, který se v dopisech vyskytuje. Musíme zde mít ovšem stále na paměti, že kolokace pracují s pravidelným opakováním, a tudíž, chceme-li zjistit každý výskyt každého slova poblíž výrazu „king“, je vhodné přenastavit parametry vyhledávání či používat pouze konkordance.

N-gramy

Dalším způsobem vyhledávání v korpusu, který může být užitečný i při historickém bádání, je tvoření tzv. n-gramů. Jedná se prakticky o posloupnost určitých částí textu, v našem případě slov, která tím, že se v textu vyskytují spolu, tvoří bigramy, trigramy a podobně (v závislosti na tom, kolik po sobě jdoucích slov sledujeme). V praxi s nimi pracujeme tím způsobem, že pouze v programu navolíme počet n-gramů, načež se zobrazí jejich seznam seřazený podle četnosti (pro nás trigramy s frekvencí opakování 30):

	Word	Count
1	to your I	202
2	to her majestie	117
3	that your I	75
4	take mi leve	72
5	plese your I	70
6	may plese your	63
7	that her majestie	58
8	honorable good I	55
9	it may plese	51
10	her majestie to	49

	Word	Count
11	of her majesties	46
12	of her majestie	46
13	your good I	44
14	prince of orenge	43
15	that i may	41
16	i take mi	41
17	right honorable good	34
18	that i am	32
19	of your I	32

Jak si můžeme všimnout, většinu z tohoto seznamu tvoří pravděpodobně zdvořilostní fráze; kdybychom tedy přistupovali takto k souboru pramenů, jejichž povahu ani obsah neznáme, můžeme jen z tohoto usuzovat, že se bude jednat s největší pravděpodobností o korespondenci týkající se státních záležitostí. Zajímavý je zde nicméně i výskyt spojení „prince of orenge“ s frekvencí 43, z čehož se dá usuzovat, že Vilém Oranžský byl rozhodně velmi častým tématem Herleových dopisů. Pokud bychom snížili počet opakování, tabulka těchto n-gramů by byla pochopitelně značně rozsáhlejší, takový příklad (byť také v poněkud zkrácené podobě) přikládáme do přílohy č.2.

Vyhledávání n-gramů se od kolokací liší především tím, že je lze používat pro celý text. Na rozdíl od předchozího kroku, kdy jsme vyhledávali určitý výraz, tedy „king“, nezadáváme při tomto hledání žádné další parametry, než je počet slov, která chceme zobrazit a frekvenci, s jakou se objevují v textu. Program Sketch Engine umožňuje nicméně pouze vyhledávání slov stojících bezprostředně vedle sebe, není zde možnost vyhledat n-gramy s vynecháním určitých částí, tzv. skip-n-gramy.

Klíčová slova

Poslední možností vyhledávání v korpusu, již v této kapitole zmíníme, je hledání pomocí klíčových slov. Jedná se o metodu, která je do jisté míry podobná metodě tematické koncentrace. Účelem je zde hledání slov, která vystihují tematické zaměření textu, ale na rozdíl od klasické tematické koncentrace nevychází tato metoda z h-bodu, nýbrž porovnává relativní frekvence ve zkoumaném textu s tzv. referenčním korpusem, což je předpřipravená databáze, která v by v ideální případě měla být co největší, aby mohla sloužit jako reprezentativní model jazyka. Často se pro tyto účely používají tzv.

národní korpusy čili korpusy obsahující co největší počet textů konkrétního jazyka – v našem případě jsme ale použili korpus historických tisků v angličtině 15.-19. století, jelikož ze všech možných referenčních korpusů, které Sketch Engine nabízí, byl tento korpus k Herleho korespondenci časově a obsahově nejbližší. Do výsledku se potom dostávají slova, která mají výrazně větší relativní frekvenci než v referenčním korpusu.

Výsledek vypadá následujícím způsobem:

	Word		Word		Word		Word		Word
1	nott	11	whatt	21	cowd	31	theme	41	yow
2	majestie	12	owtt	22	humblye	32	frynd	42	som
3	lre	13	att	23	enmye	33	leve	43	Yett
4	butt	14	cawse	24	Cowncell	34	mi	44	Receved
5	wolld	15	yett	25	bettwen	35	Nott	45	monsieur
6	uppon	16	ani	26	plese	36	Erlle	46	cowntenance
7	ytt	17	beffore	27	secrett	37	abowtt	47	howse
8	sholld	18	verey	28	servyce	38	Herlleli	48	Contrey
9	grett	19	Grett	29	frawnce	39	frenche	49	towching
10	sellf	20	allso	30	mene	40	havyng	50	dutye

Jak je zde patrné, dostala se do výsledného seznamu především slova, která se vyznačují nekonvenční pravopisnou podobou. Nejfrekventovanější slova jsou tedy synsémantika psaná jinak, než je zvykem v referenčním korpusu, nicméně na nižších příčkách můžeme sledovat i výrazy jako „Conwncell“ nebo „frenche“, což nám zase může prozradit něco o tématech Herleových dopisů.

Korpusová metoda je tedy schopna historii nabídnout možnost základního zpracování velkého množství pramenů, z nichž později budou vybrány pro důkladnější kvalitativní analýzu pouze některé, a to na základě zvolených parametrů (např. texty zaměřené na jisté téma nebo ty, v nichž se zmiňuje jistá historická osobnost nebo událost). Jinou možností by pak mohlo být využití v interdisciplinárním historicko-lingvistickém bádání v duchu Foucaultovy archeologie vědění, které by se v korpusu pramenů zaměřilo na způsob používání (zejména s ohledem na kontext) určitých slov vystihujících zvolené téma nebo diskurz. Tak například v korpusu Herleho korespondence bychom mohli zkoumat, jakým způsobem autor mluví o francouzském nebo španělském králi v porovnání s anglickou královnou (a jaký tudíž má vůči těmto vládcům vztah).

Metody strojového učení, určování autorství

Co přesně se rozumí pod pojmem určování autorství je zřejmě patrné z názvu – jde o soubor metod a přístupů, jakými lze identifikovat autora konkrétního textu. Základní předpoklad této metody spočívá v tom, že každý člověk má vlastní unikátní autorský styl, který se dá detekovat pomocí jistých statistiky měřitelných vlastností textu.[207] Proto se tato metoda rovněž nazývá stylometrie, a, jak už jsme zmiňovali v předchozí části, je pod tímto názvem známá již od 19. století.

Za zakladatele stylometrie se tradičně považuje Thomas Mendenhall, který na přelomu 19. a 20. století publikoval několik studií, v nichž se na základě průměrné délky slova pokusil odlišit Shakespearovy texty od textů psaných Francisem Baconem a několika dalšími autory.[208] Nejznámější je ale práce Fredericka Mostellera a Davida Wallace pod názvem *Inference and Disputed Authorship: The Federalist*, která se zabývala, jak plyne z názvu, tzv. Listy Federalistů. Jak už jsme zmiňovali, jedná se o sérii politických propagačních textů na podporu americké ústavy, které byly pod pseudonymem Publius vydány v letech 1787-1788.[209] Skuteční autoři většiny z těchto textů jsou již dlouho známi, pouze pro dvanáct z nich neexistuje žádný záznam o původci textu. Většina badatelů se shodla na tom, že se jednalo buď o Hamiltona, anebo o Madisona, ale definitivní svědectví poskytla teprve stylometrická analýza, která s pravděpodobností 97 procent přiřkla autorství Madisonovi.[210]

Určování autorství se dá používat pro mnoho různých oborů, je využitelné například v literární vědě, forenzní lingvistice či dokonce při práci tajných služeb. V současné době se stylometrie rovněž používá pro zjištění určitých sociálních vlastností autora (pohlaví, věk, rodný jazyk atd.) místo jeho samotné identifikace.[211] Význam určování autorství pro historické účely je poměrně zřejmý – metody určování autorství nám mohou pomoci vytipovat možného autora konkrétního textu, pokud jen nám neznámý, nebo pokud se lze domnívat, že text nebyl psán tím autorem, kterému je

[207] JUOLA, P. Authorship Attribution, s. 7.
[208] MARTINDALE, C., MCKENZIE D. On the Utility of Content Analysis in Authorship Attribution: The Federalist., 259.
[209] FORSYTH, R. Stylistic Structures: A Computational Approach to Text Classification, s. 9-10.
[210] MARTINDALE, C., MCKENZIE D. On the Utility of Content Analysis in Authorship Attribution: The Federalist., 268.
[211] STAMATATOS, E. A Survey Of Modern Authorship Attribution Methods, s. 2-3.

připisován. Nejlépe využitelné jsou tyto metody při práci se souborem pramenů podobného typu, z nichž někteří autoři jsou známí a zbylí ne, přičemž existuje hypotéza, že autory neidentifikovaných textů jsou autoři zbývajících textů ze souboru, kteří z nějakého důvodu nejsou uvedeni. Toto je mezi jinými i případ zmiňovaného souboru Listů Federalistů.

Pro potřeby naší práce používání opět soubor dopisů z Herleho korespondence v podobě databáze, která byla excerpována z internetových stránek. Tyto dopisy jsou až na dvě výjimky jménem autora označené, nicméně pro představení a otestování funkčnosti metody budeme k tomuto souboru přistupovat, jako bychom jejich autory neznali. Dopisy, které nemají určeného pisatele, budou otestovány spolu s ostatními, přičemž naše „cvičná" otázka zde může znít, zdali jsme schopni určit, kdo je právě jejich autorem.

Samotné určování autorství jako takové může být založeno na různých parametrech textu – celkový počet možností se pohybuje v současné době kolem tisíce, nicméně většina a z nich vychází z takových statistických vlastností jako průměrná délka věty nebo průměrná délka slova, používání interpunkce, jako je počet čárek, středníků a dalších interpunkčních znamének v textu, či na základě frekvence používání určitých, především neplnovýznamových, slov.[212] V našem případě byly použity tyto tři metody: první je kombinace průměrné délky věty, průměrné délky slova a relativní frekvence čárek a teček v textu (na sto slov), druhou frekvence výskytu písmen v textu a třetí je porovnávání frekvence neplnovýznamových slov (synsémantik), v našem případě seznamu padesáti spojek, předložek a tvarů pomocných sloves (např. the, to, of, in, for, a, be, with…). Studie věnované určovaní autorství zpravidla vychází z delších seznamů: ačkoliv pro práci s Listy Federalistů stačilo pouhých 30 slov, současní autoři zakládají svoje měření na 150, 300, nebo dokonce 675 slovech.[213] My jsme nicméně z důvodu lepší názornosti a přehlednosti zvolili nižší počet.

Při samotném postupu byly v Pythonu v rámci skriptu vytvořeny tři tabulky, přičemž v každé z nich byly v řádcích konkrétní dopisy a ve sloupcích měřitelné vlastnosti – čili například při měření frekvence synsémantik (třetí tabulka) obsahoval

[212] Tamtéž, s. 1.
[213] STAMATATOS, E. A Survey Of Modern Authorship Attribution Methods, s. 5.

každý jednotlivý sloupec frekvenci každého jednoho konkrétního slova („the“ nebo „of“ atd.) pro každý konkrétní text; v tomto případě se jednalo o padesát sloupců.

Tabulky byly vytvořeny následujícím skriptem za pomoci knihovny Pandas:

```
def freq_char(text, char):
    freq = text.count(char)
    length = len(my_lib.tokenize_regex(text))
    return round(freq / (length / 100), 2)

def freq_word(text, word):
    tokens = my_lib.tokenize_regex(text)
    freq = tokens.count(word)
    length = len(tokens)
    return round(freq / (length / 100), 2)

def sentence_len(text):
    length = len(my_lib.tokenize_regex(text))
    freq = text.count('.')
    try:
        output = round(length / freq)
    except:
        output = 0
    return output

def token_len(text):
    return round(len(text) / len(my_lib.tokenize_regex(text)), 2)

with open('synsem_list.txt') as input_file:
    synsem_list = input_file.read().split('\n')
with open("herle.json") as my_file:
    data = my_file.read()
    herle_dict = json.loads(data)

table1 = []
table2 = []
table3 = []
abc = list('abcdefghijklmnopqrstxyz')

for item in herle_dict:
    table1.append ([freq_char(item['text'], ','),
    freq_char(item['text'], '.'),
    sentence_len(item['text']),
    token_len(item['text'])])
    table2_row = []
    table3_row = []
    for word in synsem_list:
        table2_row.append(freq_word(item['text'], word))
```

```
    table2.append(table2_row)
    for letter in abc:
        table3_row.append(item['text'].count(letter) / len(item['text']))
    table3.append(table3_row)

df1 = pd.DataFrame(table1)
df1.columns = ['comma', 'period', 'sentL', 'wordL']
df2 = pd.DataFrame(table3)
df2.columns = abc
df3 = pd.DataFrame(table2)
df3.columns = synsem_list
```

Na začátku jsou uvedeny jednotlivé funkce pro počítání jednotlivých vlastností (frekvence určitého písmene nebo znaku v textu, průměrná délka slova, věty…), následně byl pomocí příkazu „with“ načten seznam zkoumaných pomocných slov uložený v samostatném textovém souboru (vytvořen ručně) a poté opět pomocí „with“ načtený uložený .json soubor s texty. Jednotlivé veličiny vyjadřující vlastnosti a frekvence pro každý dopis byly uloženy do seznamů, které byly následně převedeny do Pandas tabulek. Tyto tabulky by mohly být v případě potřeby zobrazeny nebo uloženy – v našem skriptu jsme se tomuto kroku ovšem vyhnuli.

Následně byly na takto získaná data aplikovány metody strojového učení – napřed unsupervised a pak supervised. Strojové učení zjednodušeně znamená způsob automatického hledání tříd (podskupin vytvořených na základě vlastností souboru, např. frekvence jednotlivých synsémantik). Pochopitelně tento model pracuje jen na základě čísel v tabulce, kterou potřebujeme dopředu připravit ve formátu, jenž bude pro počítač zpracovatelný – nelze například pracovat s tabulkou, ve které by byla textová data; vždy je nutné vymyslet způsob, jak data převést do číselné podoby. Jádro přístupu poté spočívá v tom, že tento námi vytvořený model (opět v podobě skriptu v Pythonu) umí propojit vlastnosti z různých sloupců tabulky a na základě všech těchto vlastností dohromady rozhodne, do které konkrétní podskupiny patří daná položka, v našem případě konkrétní dopis. Jinými slovy, systém prohlédne tabulku a všechny texty rozdělí na skupiny na základě toho, jaké vlastnosti tyto texty mají.

V současné době existuje poměrně velké množství knihoven v Pythonu, které jsou právě ke strojovému učení určeny, přičemž mají všechny statistické postupy zabudované v sobě a vyžadují jen minimální účast uživatele, jehož jediný úkol spočívá v tom, aby

jednotlivé funkce vyvolával. V rámci této práce jsme pracovali především s knihovnou scikit-learn.

Strojové učení metodou supervised spočívá v tom, že uživatel napřed připraví tabulku (tzv. trénovací dataset), ve které je explicitně vyjádřená příslušnost jednotlivých položek ke konkrétní skupině. Jinak řečeno, počítači se „ukáže" jakým způsobem jsou data v tomto trénovacím datasetu rozdělená do konkrétních tříd. [214] Následně se na tomto datasetu model tzv. natrénuje čili se naučí spojovat jednotlivé vlastnosti (frekvence synsémantik) se zadanou třídou (v našem případě se jednalo o dvě třídy – Herle, označovaný číslem 0, a ostatní autoři, označovaní číslem 1). Poté, co se model natrénuje, aby rozeznával tyto dvě třídy, se použije druhý dataset, tzv. testovací, v němž už model nemá přístup k informaci o tom, do které skupiny položky patří. Jinými slovy, na základě předchozích „znalostí" z trénovacího datasetu by se měl model pokusit uhodnout příslušnost ke třídám i u položek v datasetu testovacím, u nichž neví, ke které třídě patří. Prakticky se tedy na předchozích datech naučí, jak rozdělit data nová.

Unsupervised naproti tomu nevyžaduje účast uživatele a přípravu trénovacího datasetu, jelikož se model v podstatě nic neučí dopředu, pouze se snaží rozdělit na skupiny dataset, který má k dispozici, a to buď na počet skupin určených uživatelem nebo na dopředu nespecifikovaný počet, které na základě vlastností v tabulce model odvodí sám. V případě supervised tedy uživatel ukazuje modelu, jak má vypadat rozdělení na skupiny a tento to poté pouze napodobuje, zatímco v případě unsupervised learning základ pro rozdělení datasetu na třídy vytváří model sám, nejčastěji na základě toho, jak moc jsou vlastnosti konkrétních položek od sebe vzdálené nebo sobě naopak podobné. [215]

Předtím, než tabulky použijeme jako vstup pro model, je potřeba provést několik dalších kroků: prvním z nich je převod do podoby matice. Matice je velmi jednoduše řečeno seznam seznamů čísel – jedná se tedy o seznam položek, z nichž každá je tvořena seznamem čísel. Matice jsou jediným formátem dat, se kterým umí strojové učení pracovat. Dalším krokem je pak tzv. škálování čili převedení čísel do formátu, který eliminuje rozdíl rozsahu jednotlivých hodnot (například je jasné, že slova „a" nebo „the" budou mít výrazně větší frekvenci než „without"; my ale potřebujeme, aby každý ze sloupců v tabulce měl pro model stejnou váhu). V praxi to poté znamená, že všechna čísla

[214] MOHRI, M., ROSTAMIZADEH, A., TALWAKAR, A. Foundations of Machine Learning, s. 7.
[215] Tamtéž.

se převedou do podoby reálných čísel v rozsahu mezi -1 a 1. Průměr všech hodnot ve sloupci je poté vyjádřený 0, nejvyšší hodnota 1, nejnižší naopak -1.

Tyto úkony provedeme jednoduše následujícím způsobem:

```
array = df3.to_numpy()
array = scale(array, axis=0)
```

Následně na takto získaná data použijeme hierarchické shlukování, tzv. clustrování. Jedná se o jednu z metod unsupervised strojového učení, přičemž základním smyslem tohoto přístupu je postupné vytvoření shluků složených ze spojení nejbližších bodů, kdy jeden bod reprezentuje jednu položku v seznamu, v našem případě jeden dopis. Největší shluky poté tvoří třídy. [216] Pro naše účely jsme použili model, který v datasetu hledá předem určený počet tříd, v tomto případě dvě – Herle (0) a ostatní (1). Je důležité si uvědomit, že model tehdy ví pouze to, že musí hledat dvě třídy, ale nemá jediné vodítko ohledně toho, na základě čeho by třídění mělo vypadat.

V tomto se ukazuje nebezpečí tohoto přístupu, protože model, který dostal na vstupu první tabulku (průměrná délka slova, věty atd.) a rovněž model, který dostal druhou (frekvence písmen) sice detekoval v datasetu dvě třídy, ale do druhé třídy zařadil v obou případech pouze jeden text z celkového počtu 303. Toto pochopitelně nejsou výsledky, které se dají použít, a můžeme tedy tvrdit, že minimálně na souboru takovéhoto typu korespondence se tyto dvě metody určení autorství neosvědčily. Důvodem může být jak neadekvátnost samotné metody, tak i například nedostatečná velikost textů nebo nejednoznačnost přepisu dopisu, s kterýmžto problémem se ovšem bohužel bude bádání v oblasti historie muset setkávat.

Proto je lepší zaměřit se na metodu založenou na měření frekvence vybraných neplnovýznamových slov. Tento model byl napřed vyvolán ve skriptu tímto příkazem:

```
cluster = AgglomerativeClustering(n_clusters=2, affinity='manhattan',
linkage='complete')
cluster.fit_predict(array)
```

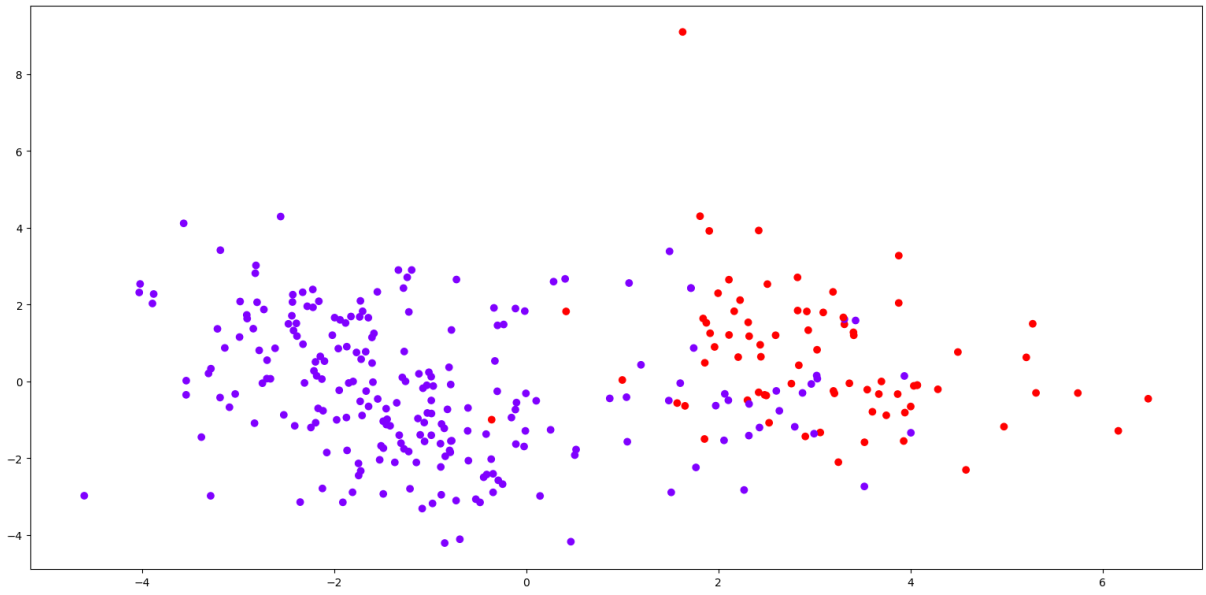
Ještě předtím, než budeme schopni pomocí clustrování vykreslit vizualizaci takto získaných dat, musíme provést tzv. Principal Component Analysis, též PCA čili analýzu hlavních komponent. Tento krok jednoduše spojí vlastnosti (našich 50 sloupců) do dvou

[216] THEODORIDIS, S., KOUTROUMBAS, K. Pattern Recognition, s. 397-401.

– jednu použijeme v grafu na osu x, druhou na osu y. Ve výsledku tedy získáme dvě syntetické vlastnosti, pomocí kterých budeme postupovat dál. Není nutné tento krok provádět ručně, stačí následující zápis:

```
pca = PCA(n_components=2)
pca.fit(array)
data2D = pca.transform(array)
plot_x = data2D[:,0]
plot_y = data2D[:,1]
```

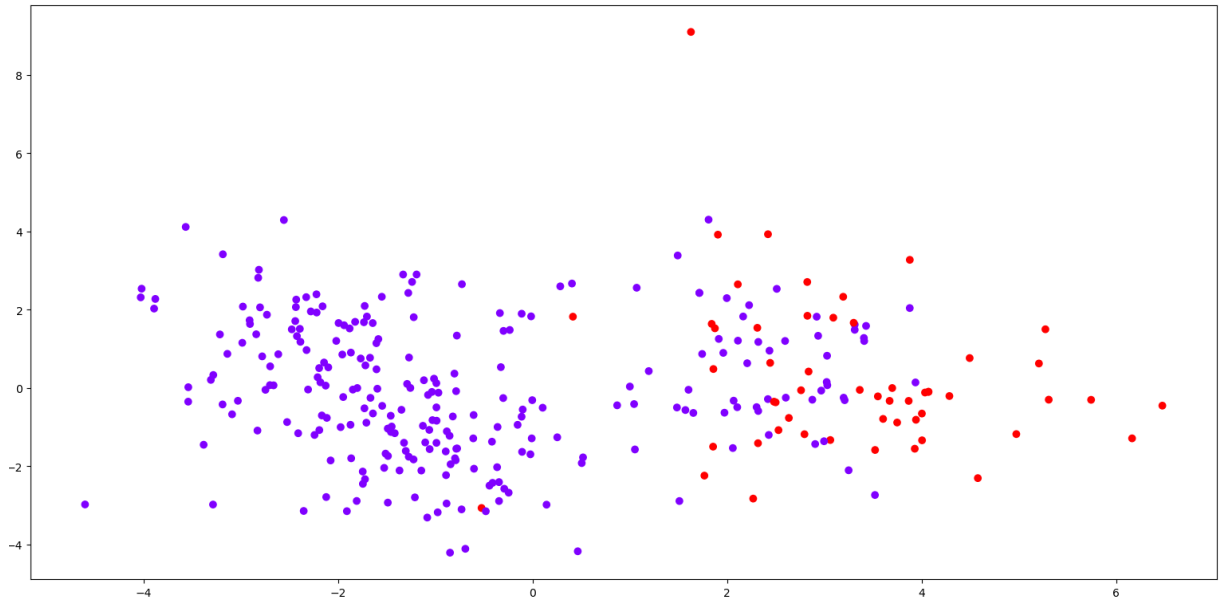
V následujícím kroku už jen pomocí hierarchického shlukování necháme vykreslit vizualizaci, přičemž je patrné, že model dokázal poměrně dobře rozdělit texty na dvě hledané třídy. Výsledný graf vypadal takto:



Každý z jednotlivých bodů označuje jeden text, přičemž si můžeme všimnout, že modré body, které znázorňují texty, které podle modelu napsal Herle, jsou většinově umístěny v jedné straně grafu, kdežto červené, jež mají mít podle modelu za autory někoho jiného, jsou více rozptýlené a je jich méně. Bod, který se vyskytuje v horní části grafu, mimo ostatní shluky, se tradičně nazývá outlier, neboli extrémní hodnota. V tomto případě se jedná o text s abnormálně krátkým obsahem (jedna věta), což je jednou z nejčastějších příčin extrémních hodnot při zkoumání textů obecně. Uvádí se, že pro

nejlepší fungování stylometrických metod je potřeba, aby každý ze zkoumaných textů měl délku aspoň 250 slov.[217]

Když porovnáme takto získané výsledky se skutečnými autory našich textů, jak je zaznamenali autoři projektu Letters of William Herle, vypadá situace takto:



Je patrné, že v pravém shluku došlo při predikování k chybě, kdy model očekával větší pravidelnost, a tudíž označil texty, které reálně Herle psal, za texty, jejichž autorem je někdo jiný. Stejně jako u většiny jiných měření podobných tomuto je i zde problematická především jakási hraniční část, kdy model není schopen s jistotou obě skupiny odlišit. Je zřejmé, že čím víc se blížíme k pravé části grafu, tím méně chyb model dělá. Většina chyb se nachází v poměrně hustém shluku nad hranicí bodu 2 na ose x. Co nám to tedy říká o této metodě?

Výsledná úspěšnost se dle komparace výsledků měření s reálnými daty pohybovala mezi 79 % a 88 % v závislosti na nastavení různých drobných parametrů modelu (typ vzdálenosti, způsobu propojování bodů do shluků atd.). Model se obecně ukazuje být úspěšnější v odhalování dopisů, jejich autorem byl Herle, nicméně je to zanedbatelný rozdíl – 88 % oproti 87 %. Kdybychom to převedli na reálný počet chyb, odhalil model jeho autorství v 220 případech z 249. Pokud jde o texty, které Herle nenapsal, predikoval model správně ve 47 případech z 54. Na základě toho můžeme usuzovat, že se zvolená metoda určování autorství v kombinaci s hierarchickým

[217] FORSYTH, R. Feature Finding for Text Classification, s. 5.

shlukováním poměrně dobře osvědčila a pokud bychom měli k dispozici obdobný soubor textů, o němž bychom věděli, že většinu z těchto textů napsal jeden autor a nějaké ne (nevíme přesně které), mohli bychom texty v tomto souboru na základě této metody relativně úspěšně klasifikovat.

Co se týče dvou neznámých textů, o nichž jsme hovořili na začátku kapitoly, podle predikce našeho modelu vypadá, že se skutečně nejedná o texty z pera Herleho, nicméně vzhledem k malému počtu textů ostatních autorů, nejsme schopni s jistotou říci, který z nich by mohl být autorem těchto dvou dopisů.

Supervised learning, jak už jsme řekli, spočívá v tom, že dopředu určíme, jaké položky do které třídy patří, na základě čehož by se měl být model schopen naučit rozeznávat zadané třídy i na neznámých datech. Pro naše účely byla použita metoda SVM jinak také metoda podpůrných vektorů, což je nejjednodušší metoda supervised učení. Jedná se o přístup, kdy se model pokusí oddělit data tím, že vytvoří „přímku“ rozdělující dataset na dvě části. [218] Pokud bychom se dívali na přechozí graf, můžeme si představit, že by model nakreslil mezi ně rovnou čáru, kterou by se pokusil oddělit oba dva shluky. Tuto čáru by model určil tak, aby body z každé její strany patřily pouze do jedné třídy podle toho, jak jsou určeny v testovacím datasetu.

Před samotným trénováním tohoto typu modelu je třeba rozdělit dataset na trénovací a testovací, přičemž by se každý z nich měl skládat ze dvou různých částí (tabulek); v tabulce x budou uloženy hodnoty, na jejichž základě by měl model odvozovat třídy a v tabulce y pak samotné přiřazení položek ke třídám (v našem případě 1 nebo 0 = autor není Herle nebo je Herle). Při trénovaní se tedy model pokusí spojit data z tabulky x s výsledkem v tabulce y. Bude tedy hledat takové rozložení čísel, které odpovídá každé ze tříd.

Dataset, který se v našem případě skládal celkem z 303 textů, z nich 54 nepsal Herle, tedy rozdělíme pomocí funkce z knihovny sklearn a následně vyvoláme trénování modelu:

```
def herle (author):
    if 'Herle' in author: return 0
    else: return 1

x = array
```

[218] FELDMAN, R., SANGER, J. The Text Mining Handbook, s. 76.

```
y = np.array([herle(x['author']) for x in herle_dict])
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33,
random_state=42)
model = SVC(gamma='auto')
model.fit(x_train, y_train)
```

První funkce v této části skriptu má za účel převedení textového typu dat z položky obsahující záznam o autorovi dopisu do binární číselné podoby. Následující část je poté samotné rozdělení na datasety a trénování.

Po trénování modelu byla otestována jeho úspěšnost na testovacím datasetu, což je ta část dat, kterou model „nezná“ a nemá přístup k výsledkům a pokouší se na ní aplikovat ty postupy, které se „naučil“ na trénovacím datasetu. V našem případě se úspěšnost takto navolenému modelu rovnala 88 %, což je stejný výsledek, kterého jsme dosáhli i při použití metody hierarchického shlukování. Je důležité si ale uvědomit, že trénovací dataset, který jsme předkládali modelu, nebyl v tomto případě vyvážený; to znamená, že textů jedné třídy tam bylo výrazně více než textů druhé třídy (textů, jejichž autorem je Herle, bylo výrazně více než textů jiných autorů), což mohlo ovlivnit výkonost modelu.

Tato hypotéza se ukázala být správnou, jelikož po přípravě vyváženého datasetu, který obsahoval 50 textů Herleho a 50 textů jiných autorů, se přesnost modelu na testovacích datech zvýšila na 93 %. Můžeme tedy tvrdit, že při správné přípravě trénovacího datasetu se zvolená metoda určování autorství ukázala být dostatečně spolehlivou i přes svou jednoduchost.[219]

Praktické použití takto natrénovaného SVM modelu by mohlo spočívat kupříkladu v tom, že v případě, kdy bychom objevili další text, například ze stejného období, o kterém bychom se měli důvod domnívat, že jeho autorem je William Herle, mohli bychom s pravděpodobností 93 % určit, zdali je tomu skutečně tak. Je ovšem důležité si uvědomit, že zjištění získaná pomocí metod strojového učení budou mít vždy pouze pravděpodobnostní charakter, nikdy tedy nemůžeme mít absolutní jistotu, minimálně bez přihlédnutí k dalším faktorům a doplněním metody o jiné přístupy včetně

[219] Vyšší úspěšnosti bychom mohli dosáhnout používáním většího a vyváženějšího datasetu, což ovšem v tomto případě – stejně jako u mnoha dalších historických projektů – není ze zřejmých důvodů možné. Dalším řešením by mohlo být poté zvýšení počtu používaných frekvencí nebo používání jiných metod strojového učení, např. neuronových sítí, které jsou díky pokročilé technologii schopny efektivněji pracovat na zadaném datasetu.

kvalitativních. Na druhou stranu, ani ručně zpracovaná stylometrická analýza nezaručuje pochopitelně absolutní přesnost, nicméně není pochopitelně možné procentuálně vyjádřit chybovost manuálního zpracování člověkem.

Ačkoliv fungování metody strojového určení bylo v této kapitole znázorněno ve spojení s určováním autorství, není to její jediné možné využití v rámci historie. Stejně jako ve spoustě dalších přírodovědných a humanitních oborů, i v historickém bádání lze používat hierarchické shlukování a jiné metody unsupervised strojového učení i k formulování hypotéz a výzkumných otázek pro zkoumání většího souboru počítačově zpracovatelných pramenů, které není možné kvůli rozsahu důkladně prozkoumat tradičním způsobem. Nástroje a metody pro tento krok může v současné době poskytnout nové interdisciplinární odvětví pod názvem „data mining“, které spojuje statistiku, informatiku a strojové učení. Hlavním cílem data miningu je odhalení skrytých vzorců a souvislostí ve větších, předem neprozkoumaných souborech dat za pomocí komputačních nástrojů. Tradiční hypoteticko-deduktivní přístup k vědě se při postupech data miningu obrací takříkajíc vzhůru nohama: badatel přistupuje v tomto případě k datům jako k „tabula rasa“, tj. bez předchozích hypotéz a předpokladů, nechává naopak hypotézu „vyvstat“ ze samotných dat na základě odhalené struktury datasetu a pravidelností v ní.[220]

[220] LAROSE, D. Discovering Knowledge in Data: Introduction to Data Mining, s. 2-3.

Závěr

V této práci jsme na příkladu vývoje této disciplíny ukázali, že digital humanities je moderní, dynamicky se rozvíjející obor, který zakládá na využití počítačových metod, webových technologií, internetu a multimediálních zdrojů. Důležitou složkou tohoto oboru je digitální historie, což je nový interdisciplinární směr v oblasti historického bádání, jenž vznikl v 90. letech na vlně rozvoje internetu a digitálních sbírek pramenů. Tento obor přesto nebyl ničím radikálně novým, nýbrž navazoval zároveň na anglofonní tradici tzv. history computing z 50. let a zároveň na historickou informační vědu, vznikající v evropském prostředí v 80. letech. Ukázali jsme, že vývoj používání počítačových metod v historii neprobíhal stejně v Evropě a v Americe; zatímco Amerika prožila obrovskou vlnu zájmu o kvantitativní metody v podobě kliometrie, v Evropě nebyl touto dobou o metody podobného typu příliš velký zájem, a naopak se zde začaly rozvíjet v období, kdy v Americe začínaly upadat. Teprve v 90. letech po vzniku internetu se začaly tyto dva proudy propojovat do současné podoby digitální historie.

Po uvedení různých počítačových metod v teoretické části práce jsme se v praktické části zaměřili na ukázky jednotlivých postupů při digitálním zpracování historických pramenů, které vycházejí především z metod používaných tradičně v počítačové a kvantitativní lingvistice. Postupně bylo předvedeno vytěžení textu z webových stránek, vytvoření databáze, frekvenčního slovníku, analýza tematické koncentrace a korpusové zpracování vytěžených pramenů. V poslední kapitole jsme znázornili metodu strojového učení na příkladu určování autorství.

Práce ukázala užitečnost zkoumaných počítačových metod:

1) při původním zpracování pramene, který je součástí většího souboru, jenž není možné celý prozkoumat klasickým kvalitativním způsobem
2) v prvotních fázích výzkumu při formulování hypotéz a výzkumných otázek na základě statistických metod hierarchického shlukování a obecně data miningu
3) jako podpůrný nástroj při testování hypotéz, který se opírá o tradiční kvalitativní analýzu pramene a doplňuje ji

Zároveň ale nesmíme zapomínat na to, že je vždy potřeba dostupné metody upravit pro účely konkrétního historického výzkumu, abychom zajistili, že jejich používání není samoúčelné. Také je nutné vždy kvantitativní data doplnit a podložit důkladným kvalitativním výzkumem.

Seznam zdrojů

Literatura:

BERRY, D. *Introduction: Understanding the Digital Humanities.* In: BERRY, D. (ed.). Understanding Digital Humanities. New York 2012, s. 1-20.

BERRY, D., FAGERJORD, A. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge 2017.

BILANSKY, A. *Search, Reading and the Rise of Database*. In: Digital Scholarship in the Humanities, č. 3, 2017, s. 511-527.

BOONSTRA, O., BREURE, L., DOORN, P. *Past, Present and Future of Histocial Information Science*. Amsterdam 2006.

BURKE, P. *Languages and Communities in Early Modern Europe.* Cambridge 2004.

BURTON, D. M. *Automated Concordances and Word Indexes: The Early Sixties and the Early Centers*. In: Computers and the Humanities, č. 15, 1981, s. 83-100.

BURTON, D. M. *Automated Concordances and Word Indexes: The Fifties.* In: Computers and the Humanities, č. 15, 1981, s. 1-14.

BUSA, R. *Foreword: Perspectives on the Digital Humanities.* In: SCHREIBMAN, S. a kol. (eds). A Companion to Digital Humanities. Oxford 2004, s. XVI-XXI.

ČECH, R., GARABIK, R., ALTMANN, G. *Testing the Thematic Concentration of Text*. In: Journal of Quantitative Linguistics, č. 3, 2015, s. 215-232.

ČECH, R., POPESCU, I., ALTMANN, G. *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc 2014.

DASH, N. *Corpus Linguistics: An Introduction.* Delhi 2008.

DENLEY, P. *Models, Sources and Users: Historical Database Design in the 1990s*. In: History and Computing, č. 1, 1994, s. 33-43.

EITELJORG, H. *Computing for Archaeologists.* In: SCHREIBMAN, S. a kol. (eds.). A Companion to Digital Humanities. Ofxord 2004, s. 20-30.

FELDMAN, R., SANGER, J. *The Text Mining Handbook.* Cambridge 2007.

FITZPATRCIK, K. *The Humanities, Done Digitally*. In: GOLD, M. (ed.). Debates in the Digital Humanities. Minneapolis 2012, s. 12-15.

FORSYTH, R. *Feature Finding for Text Classification*. In: Literary and Linguistic Computing, č. 4, 1996, s. 163-174.

FORSYTH, R. Stylistic Structures: A Computational Approach to Text Classification. PhD Thesis. Nottingham 1996.

GREENSTEIN, D. *Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States*. In: Computers and the Humanities, č. 1, 1996, s. 351-364.

HARVEY, Ch., PRESS, J. *Databases in Historical Research*. New York 1996.

HERSHBERG, T. *The Philadelphia Social History Project: An Introduction*. In: Historical Methods Newsletter, č. 2-3, 1976.

HOCKEY, S. *SNOBOL Programming for the Humanities*. Oxford 1986.

HOCKEY, S. *The History of Humanities Computing.* In: SCHREIBMAN, S. a kol. (eds). A Companion to Digital Humanities. Oxford 2004, s. 3-19.

JUOLA, P. *Authorship Attribution*. In: Foundations and Trends in Information Retrieval, č. 3, 2006, s. 233-334.

KIRSCHEBAUM, M. *What is Digital Humanities and What's It Doing in English Departments*. In: GOLD, M. (ed.). Debates in the Digital Humanities. Minneapolis 2012, s. 3-11.

KÖHLER, R., GRZYBEK, P (eds.). *Exact Methods in the Study of Language and Text*. Berlin 2007.

LAROSE, D. *Discovering Knowledge in Data: Introduction to Data Mining*. Hoboken 2005.

MARTINDALE, C., MCKENZIE D. *On the Utility of Content Analysis in Authorship Attribution: The Federalist*. In: Computers and the Humanities, č. 4, 1995, s. 259-270.

MOHRI, M., ROSTAMIZADEH, A., TALWAKAR, A. *Foundations of Machine Learning*. Cambridge, MA 2012.

OLDERVOLL, J. *A System For Analysing Census-Type Data*. In: Historical Social Research, č. 3, 1988, s. 17-22.

POPESCU, I. (ed.) *Word Frequency Studies*. Berlin 2009.

POPESCU, I. *Text Ranking by the Weight of Highly Frequent Words*. In: GRZYBEK, P., KÖHLER, R. (eds.) Exact Methods in the Study of Language and Text, s. 555-566.

SEEFELDT, D., THOMAS, W.G. *What is Digital History?* In: Perspectives in History, č. 5, 2009.

SPIRO, L. *„This Is Why We Fight": Defining Digital Humanities.* In: GOLD, M. (ed.). Debates in the Digital Humanities. Minneapolis 2012, s. 16-35.

STAMATATOS, E. *A Survey Of Modern Authorship Attribution Methods*. In: Journal of the American Society for Information Science and Technology, č. 3, 2009, s. 538-556.

ŠTINDLOVÁ, J., MACHÁČKOVÁ E. *Texty Slezských písní Petra Bezruče prověřovány stroji*. In: Slovo a slovesnost, č. 2, 1970, s. 161-166.

THEODORIDIS, S., KOUTROUMBAS, K. *Pattern Recognition*. San Diego 2008.

THOMAS, W.G. *Computing and the Historical Imagination.* In: SCHREIBMAN, S. a kol. (eds). A Companion to Digital Humanities. Oxford 2004, s. 56-68.

Digitální zdroje:

Computing Our Future: Computer Programming and Coding. Priorities, School Curricula and Initiatives Across Europe. Dostupné z: http://www.eun.org/documents/411753/817341/Computing+our+future_final_2015.pdf [cit. 17. 08. 2019]

CVRČEK, V. *Frekvence*. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy – Nový encyklopedický slovník češtiny. 2017. Dostupné z: https://www.czechency.org/slovnik/FREKVENCE [cit. 17. 08. 2019]

East India Company. Dostupné z: http://www.eastindiacompany.amdigital.co.uk [cit. 17. 08. 2019]

George Washington Papers Collection. Dostupné z: https://www.loc.gov/collections/george-washington-papers/ [cit. 17. 08. 2019]

ILIESI Institute Lessico Intellettuale Europeo e Storia delle Idee. Dostupné z: http://www.iliesi.cnr.it/EN/index.shtml [cit. 17. 08. 2019]

Letters of William Herle Project. Dostupné z: http://www.livesandletters.ac.uk/herle/ [cit. 17. 08. 2019]

Making the History of 1989. The Fall of Communism in Eastern Europe. Dostupné z: http://chnm.gmu.edu/1989/ [cit. 17. 08. 2019]

Národní výzkumná a inovační strategie pro inteligentní specializaci České republiky 2020. Dostupné z: https://www.mpo.cz/assets/cz/podnikani/ris3-strategie/dokumenty/2019/1/Narodni_RIS3_strategie_aktualizace_2018.pdf [cit. 17. 08. 2019]

Perseus Digital Library. Dostupné z: http://www.perseus.tufts.edu [cit. 17. 08. 2019]

Project Gutenberg. Dostupné z: http://www.gutenberg.org/wiki/Main_Page [cit. 17. 08. 2019]

Rome Reborn. Dostupné z: https://www.romereborn.org/content/home-0 [cit. 17. 08. 2019]

Sketch Engine. Dostupné z: https://www.sketchengine.eu [cit. 17. 08. 2019]

The Centre for Editing Lives and Letters. Dostupné z: http://www.livesandletters.ac.uk [cit. 17. 08. 2019]

The Correspondence of William of Orange 1549-1584. Dostupné z: http://resources.huygens.knaw.nl/wvo/en [cit. 17. 08. 2019]

The Orlando Project. Feminist Literary History and Digital Humanities. Dostupné z: http://www.artsrn.ualberta.ca/orlando/ [cit. 17. 08. 2019]

The Proceedings of the Old Bailey. Dostupné z: https://www.oldbaileyonline.org/ [cit. 17. 08. 2019]

The Valley of the Shadow. Dostupné z: http://valley.lib.virginia.edu/VoS/usingvalley/valleystory.html [cit. 17. 08. 2019]

Wiki Českého národního korpusu. Dostupné z: https://wiki.korpus.cz/doku.php/kurz:pokrocile_dotazy [cit. 17. 08. 2019]

Resumé

Tato práce se věnuje problémům využití počítačových metod při analýze historického pramene v duchu tzv. digital history. První část práce, která je spíše historiografického zaměření, podává definici disciplíny digital humanities a stručně shrnuje její historický vývoj od 50. let 20. století až po současnost; zvláštní zřetel je poté brán na vývoj používání počítačových metod v historii s přihlédnutím k teoretickým a praktickým problémům, které jejich používání doprovází. Druhá část práce je pak praktickou ukázkou aplikace konkrétních komputačních metod spadajících tradičně do oblasti počítačové, korpusové a kvantitativní lingvistiky na souboru vybraných pramenů, kterými je soupis korespondence diplomata alžbětinské doby Williama Herleho. Postupně jsou zde prezentovány metody web scrapingu, tvoření databází, využití frekvenčních slovníků a tematické koncentrace, korpusové metody a strojové učení na příkladu určování autorství.

This thesis deals with the problem of application of computer methods in historical source analysis in the spirit of so called digital history. Its first part, which is historiographical, provides a definition of digital humanities and briefly summarizes historical development of the discipline from the 1950s to the present day; special attention is given to the development of computer methods in history, including the theoretical and practical problems that accompany their use. The second part of the thesis is a practical demonstration of the application of particular computational methods, that traditionally fall within the area of computer, corpus and quantitative linguistics, on a set of selected sources, which is a set of correspondence of William Herle, an English diplomat from the Elisabethan era. Presented methods are: web scraping, creation of databases, use of frequency dictionaries and thematic concentration, corpus methods and machine learning on the example of authorship attribution.

Přílohy

Příloha č. 1

Rank	Freq	Rel Freq	Type
1	14712	5.493	the
2	10833	4.045	to
3	10328	3.856	of
4	5881	2.196	that
5	4023	1.502	in
6	3851	1.438	i
7	3225	1.204	for
8	3013	1.125	a
9	2832	1.057	be
10	2706	1.010	with
11	2656	0.992	he
12	2571	0.960	by
13	2512	0.938	your
14	2507	0.936	and
15	2443	0.912	is
16	2434	0.909	as
17	2287	0.854	which
18	2208	0.824	his
19	1917	0.716	have
20	1688	0.630	l
21	1643	0.613	this
22	1573	0.587	me
23	1539	0.575	her
24	1499	0.560	they
25	1485	0.554	it
26	1471	0.549	mi
27	1374	0.513	was
28	1321	0.493	so
29	1140	0.426	had
31	1101	0.411	their
32	1041	0.389	all
33	1000	0.373	from
34	988	0.369	or
35	948	0.354	butt
36	932	0.348	do
37	910	0.340	good
38	886	0.331	on
39	871	0.325	may
40	855	0.319	majestie
41	851	0.318	hym
42	824	0.308	yow
43	794	0.296	my

44	794	0.296	att
45	782	0.292	will
46	780	0.291	hath
47	763	0.285	were
48	746	0.279	him
49	738	0.276	nott
50	726	0.271	fol
51	690	0.258	sayd
52	689	0.257	som
53	686	0.256	not
54	675	0.252	more
55	630	0.235	theme
56	627	0.234	but
57	624	0.233	other
58	592	0.221	here
59	587	0.219	there
60	580	0.217	shall
61	554	0.207	uppon
62	537	0.201	unto
63	534	0.199	master
64	525	0.196	who
65	513	0.192	most
66	509	0.190	yn
67	507	0.189	ar
68	500	0.187	them
69	488	0.182	k
70	482	0.180	ani
71	481	0.180	these
72	477	0.178	might
73	467	0.174	yf
74	465	0.174	grett
75	459	0.171	same
76	445	0.166	if
77	434	0.162	then
78	434	0.162	no
79	428	0.160	an
80	427	0.159	monsieur
81	411	0.153	at
82	406	0.152	tyme
83	405	0.151	those
84	392	0.146	well
85	391	0.146	ytt
86	386	0.144	yett
87	383	0.143	suche
88	381	0.142	you
89	376	0.140	into

90	373	0.139	made
91	370	0.138	now
92	362	0.135	bothe
93	361	0.135	wolld
94	355	0.133	than
96	334	0.125	humbly
97	332	0.124	men
98	331	0.124	th
99	330	0.123	matter
100	328	0.122	sholld
101	313	0.117	whom
102	307	0.115	p
103	304	0.114	nor
104	295	0.110	wold
105	294	0.110	herle
106	291	0.109	england
107	290	0.108	god
108	289	0.108	q
109	286	0.107	state
110	286	0.107	make
111	284	0.106	sellf
112	283	0.106	frenche
113	282	0.105	being
114	282	0.105	com
115	279	0.104	am
116	278	0.104	our
117	277	0.103	whatt
118	274	0.102	self
119	274	0.102	ye
120	270	0.101	man
121	266	0.099	majesties
122	264	0.099	things
123	260	0.097	owtt
124	258	0.096	very
125	257	0.096	take
126	257	0.096	allso
127	257	0.096	service
128	256	0.096	own
129	253	0.094	wherof
130	252	0.094	honorable
131	249	0.093	when
132	249	0.093	only
133	247	0.092	any
134	247	0.092	she
135	245	0.091	shalbe
136	244	0.091	verey

137	243	0.091	further
138	237	0.088	lre
139	235	0.088	lres
140	233	0.087	towards
141	233	0.087	where
142	232	0.087	having
143	227	0.085	whole
144	225	0.084	don
145	225	0.084	w
146	224	0.084	before
147	223	0.083	place
148	223	0.083	ij
149	223	0.083	hable
150	222	0.083	sent
151	218	0.081	honor
152	215	0.080	shold
153	214	0.080	some
154	211	0.079	how
155	211	0.079	mene
156	209	0.078	hand
157	207	0.077	favor
158	204	0.076	ether
159	203	0.076	prince
160	202	0.075	towne
161	201	0.075	present
162	201	0.075	can
163	199	0.074	yet
164	199	0.074	over
165	198	0.074	against
166	196	0.073	withall
167	196	0.073	first
168	196	0.073	without
169	194	0.072	beffore
170	193	0.072	owne
171	191	0.071	charge
172	191	0.071	cawse
173	190	0.071	rest
174	188	0.070	muche
175	187	0.070	dyd
176	186	0.069	long
177	183	0.068	end
178	182	0.068	therof

Příloha č.2

N-gram	Freq
to your l	202
to her majestie	117
that your l	75
take mi leve	72
plese your l	70
may plese your	63
that her majestie	58
honorable good l	55
it may plese	51
her majestie to	49
of her majesties	46
of her majestie	46
your good l	44
prince of orenge	43
that i may	41
i take mi	41
right honorable good	34
that i am	32
of your l	32
to her majesties	29
of this towne	28
mi right honorable	28
your honors most	27
plese yow to	26
to your honor	25
of that which	25
of his own	25
to your majestie	23
may plese yow	23
humbly i take	23
of all other	21
your honor to	20
to your good	20
of those that	20
of his owne	20
i am hable	20
her majestie of	20
tyme to tyme	19
king of spayne	19
to your honorable	18
that he may	18
take my leave	18
send your l	18
prince of parma	18

of this moneth	18
which your l	17
to whom he	17
to consyder of	17
of myne owne	17
humbly take mi	17
that i shold	16
beseche your l	16
that he wold	15
owtt of england	15
of this present	15
her majesties service	15
her majestie that	15
that he cowd	14
shalbe hable to	14
owtt of frawnce	14
of their owne	14
most humbly w	14
most humbly i	14
it to your	14
it may please	14
i take my	14
i beseche your	14
erlle of lecester	14
your good favor	13
to whom i	13
to give me	13
those of holland	13
state of things	13
plese your honor	13
mi sellf to	13
may please your	13
majestie of england	13
i humbly take	13
her majesties hands	13
that they may	12
state of these	12
please your l	12
of your majesties	12
of these contreyes	12
i desire to	12
he sayd that	12
comend to your	12
am hable to	12
all those of	12
all that i	12

william herle to	11
towards her majestie	11
to those of	11
to make his	11
sent your l	11
sayd that he	11
right gracious soveraigne	11
of your good	11
of whom i	11
of their own	11
most humblye w	11
k of spaine	11
i most humbly	11
honors most humblye	11
beseching your l	11
assuryng your l	11
all that he	11
according to your	11
your majestie to	10
verey humbly i	10
to your wisdom	10
to troble your	10
to this daye	10
to my l	10
that it may	10
of this town	10
mi good l	10
it plese your	10
i shalbe hable	10
honorable good lord	10
her majestie may	10
cowncell of estate	10