Příručka pro české uživatele programu Prospéro

(Elektronická Verze 1.0)

Tereza Klabíková Rábová, Pavel Kotlík, Simon Smith

Tato publikace vznikla za podporu Grantové agentury České republiky v rámci grantu GA16-20553S: Narativizace krize a instituce ve stranické politice (politics) a veřejných politikách (public policy).

www.narratingcrisis.com

Únor 2019

Obsah

I. Úvod

Program Prospéro je nástroj pro diskurzivní a argumentační analýzu korpusů textů, vyvinutý školou EHESS v Paříži; jeho frankofonní původ se mj. odráží v dosavadním jazyce celého uživatelského rozhraní. K Prospéru se váže web www.prosperologie.org, kde je třeba se zaregistrovat, aby bylo možno program stáhnout (dále viz oddíl Instalace). Na těchto stránkách je také fórum, kde lze v příslušných tématech vyhledat a v případě absence i doplnit popis problémů či se pro jejich řešení zapojit do diskuze uživatelů. Český překlad výběru z knihy Francise Chateauraynauda ***Prospéro. Une technologie littéraire pour les sciences humaines***, která je souhrným dílem o Prospéru, je ke stažení zde.

K čemu je Prospéro užitečný?

Prospéro umožňuje nahlédnout do korpusu textů, na jeho části i na celek – zde ukazuje hlavní aktéry textů, děje a vlastnosti, které se k nim vážou, temporalitu textů (bere v potaz jak serializaci událostí v korpusu, tak modalizaci času v argumentaci aktérů), umožňuje také práci s metadaty (kdy, kde, kdo o tématu psal). V neposlední řadě umožňuje prozkoumávat i velmi drobné detaily, ptát se konkrétně např. na jednu vazbu, slovo, spojovací výraz či např. na tři spojená substantiva, práci jednoho autora apod. – tzn. formulovat dotaz, na který lze hledat odpověď pomocí určité kombinace zástupných slov a konceptů.

Jak Prospéro indexuje texty obsažené v našem korpusu?

Prospéro prohledává texty na základě výzkumníkem definovaných/upravených jazykových slovníků (blíže viz oddíl Kategorizace slov a struktura jazykových slovníků); slova jsou primárně řazena do **čtyř hlavních skupin**: podstatná jména (*entités),* adjektiva označující vlastnosti jim přináležící či je definující (*qualités),* slovesa vyjadřující děj či akci (*épreuves)* a konečně příslovce a zároveň další argumentativní indikátory a konektory (*marqueurs)* a vedle nich ještě slova pomocná a číslovky. Shoda mezi základními pojmy Prospéra a standardními gramatickými kategoriemi je ovšem nedokonalá – slouží spíš jako prototypy.

Jaký je analytický rámec Prospéra?

Na další úrovni je obsah textů tříděn do/dle tzv. **kategorií** (sémanticky blízká seskupení *entités, épreuves, qualités* a *marqueurs*; kategorie tedy zahrnují tyto čtyři slovní druhy), **seznamů/kolekcí** (např. politických stran, médií, zemí) a **velkých postav korpusu** (umožňuje např. shlukování synonym). Rozsah a obsah těchto pojmů definuje badatel sám (blíže viz oddíl Pojmy-koncepty a pojmové slovníky: kategorie, kolekce, velké postavy), ale

zároveň navazuje na již sestavené slovníky ostatních výzkumníků. Charakteristické pro Prospéro tedy je to, že se jedná o kolektivní, neustále se rozvíjející klasifikační systém, který je možné a v dané komunitě běžné sdílet (viz stránky Prospérologie). Aktuálně dostupné pojmové slovníky jsou výsledkem práce prvních českých uživatelů a zároveň východiskem pro práci dalších.

Jak lze Prospéro využít pro analýzu textů?

Program nabízí několik dalších analytických nástrojů (diskurzivní konfigurace, vzorce (*formules*) aj.; viz oddíl Analytické nástroje Prospéra). Umožňuje vždy vidět **kontext** slov, vět, výpovědí. S pomocí dalších programů (např. Pajek nebo Gephi) je možno **výsledky vizualizovat,** vytvářet grafy, srovnávat.

Pro analýzu jakých typů textů a korpusů je Prospéro vhodný?

Prospéro vznikl v rámci výzkumné skupiny, která sociologicky zkoumá kontroverze a veřejné problémy, většinou dlouhodobé povahy, a tak je primárně určen pro typ korpusu sledující vývoj kauz v delších časových úsecích – obsahy textů se s plynutím času vyvíjejí, jejich aktéři, situace, instrumenty apod. se přidávají, mění. U jinak konstruovaných korpusů je užitečný např. pro lingvistické analýzy gramatické a pravopisné (např. i sledování chyb, velkých písmen, zkratek). Význam má rovněž pro rozbory slovotvorné a lexikální – poskytuje frekvence slovních druhů, synonym, seznamy podstatných, přídavných jmen nebo sloves (verba dicendi). Vhodný je i k popisu jevů syntaktických jako frekvence a podoby různých typů souvětí či typy podmětů, dále stylistických (typy vyjádření aj.) či diskurzivních (klíčové výrazy, ideologémy, argumentace aj.) Je zde tedy využita spíše první úroveň indexace textů a zároveň pak další analytické nástroje.

Program je vhodný jen pro **jednojazyčné texty** – při komparaci dvojjazyčných korpusů je třeba pracovat se dvěma různými verzemi Prospéra, ale pak už nelze korpusy srovnat navzájem. Některá rozhraní jsou konstruována zejména **pro texty mediální** (např. zadání metadat k textům), korpusy ale mohou obsahovat i jiné typy textů (smlouvy, předpisy, zákony). Program bere v potaz pouze text ve formátu .txt (blíže viz oddíl Příprava textů); nebere v potaz vizualitu ani celkovou podobu textu, kompozici apod., a proto není zcela vhodný pro analýzu webových stránek jako takových, neboť obsahují další znaky, obrázky, mezititulky, hypertextové odkazy apod. (většinou se užívají články z Newtonu a podobných databází).

Kolik textů může či by měl korpus obsahovat, jak mají být dlouhé?

Program je schopen zahrnout do analýzy sbírky/korpusy **desítek až tisíců** psaných, resp. tištěných textů jakékoliv délky. V takto rozsáhlých souborech program výzkumníkovi umožňuje vyhledávat poměrně rychle daný výraz/jev/koncept, a zároveň mu tím poskytuje přehled o celém korpusu. U tohoto druhu nástroje, jehož dimenze funkcionality zasahují do kvantitativního výzkumu, také platí, že větší počet textů zvyšuje pravděpodobnost zachycení dimenzí zkoumaného jevu s přiměřeným počtem výskytů. Větší korpusy lze navíc rychle a flexibilně rozdělit na menší podkorpusy.

Jak časově náročná je práce s Prospérem?

Záleží na počtu a délce textů v korpusu, na jejich formátování, ale i na způsobu analýzy podle toho, jestli provádíme spíše hrubou sondu, nebo podrobnou a vyčerpávající investigaci. Formátování textů a příprava metadat jsou úkoly různého stupně náročnosti podle míry přesnosti a obsáhlosti, kterou vyžadujeme. Záleží taky na jazyce korpusu – v případě specifického korpusu (např. obsahuje spoustu slangových výrazů nebo odborného žargonu) bude třeba rozsáhlejší úvodní přípravy textů k analýze (zadávání neznámých slov do slovníků). Náročnější příprava dat je ale vyvážena tím, že Prospéro umožňuje snadno analyzovat i obří korpusy, tedy se počáteční časová investice z dlouhodobého hlediska vyplatí.

Vzhledem k tomu, že rozhraní Prospéra je ve francouzštině, uvádíme názvy funkcí či tlačítek v původním znění *kurzívou*, v případě potřeby je pak překládáme a/nebo dále vysvětlujeme. V dohledné době bude zpřístupněna i anglická verze rozhraní.

II. Příprava programu

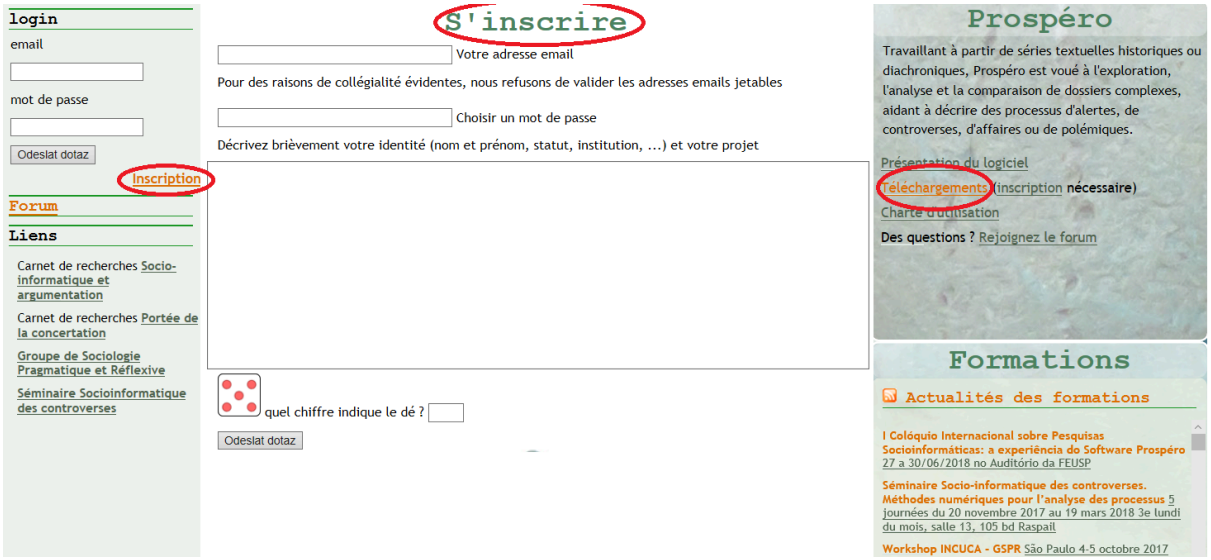
1. Instalace

Tato kapitola se zaměřuje na **registraci** a **instalaci programu Prospéro**. Představí podrobný popis celého procesu instalace včetně technických specifik a požadavků.

Následné **první spuštění** nainstalovaného programu pak provede uživatele důležitými nastaveními.

Co dělat v případě potíží?

Pokud nastane během instalace či používání programu situace, kterou nebudete schopni sami vyřešit, můžete vznést konkrétní dotaz na stránkách Prospérologie.[1] Zde naleznete diskuzní fórum, které je vhodné jak pro dotazy obecného rázu, tak i specifické problémy. Jelikož se zároveň jedná o platformu pro české uživatele Prospéra, sdílení praktických, teoretických, i analytických problémů i nápadů je vítáno přímo v české sekci.[2]



1.1 Registrace

Před samotnou instalací je třeba nejprve vyplnit registrační údaje na stránkách Prospérologie. V sekci login (levý sloupec na obrázku výše) zahájíme proces registrace přes odkaz *inscription* (registrace). V novém okně (prostřední sloupec) vyplníme následující pole:

a) e-mailovou adresu (*Votre adresse e-mail*),
b) heslo pro příští přihlášení (*Choisir un mot de passe*),
c) doplňující informace (*Décrivez brièvement votre identité (...) et votre projet*), například o akademickém zaměření, budoucím výzkumném záměru pro využití Prospéra apod.

Po vyplnění všech potřebných údajů je ještě třeba vložit číslo na kostce do volného pole (potvrdíte, že nejste robot). Následně potvrdíme stisknutím tlačítka Odeslat odkaz. Na zadanou e-mailovou adresu přijde potvrzovací zpráva.

[1] http://prosperologie.org/forum/index.php
[2] http://prosperologie.org/forum/viewforum.php?f=34

Potvrzovací zpráva (e-mail) nepřijde obratem, jelikož musí být přihlašovací údaje nejprve zpracovány, a to v rámci časových možností autorů Prospéra. Registrační údaje tak mohou přijít s odstupem několika dní.

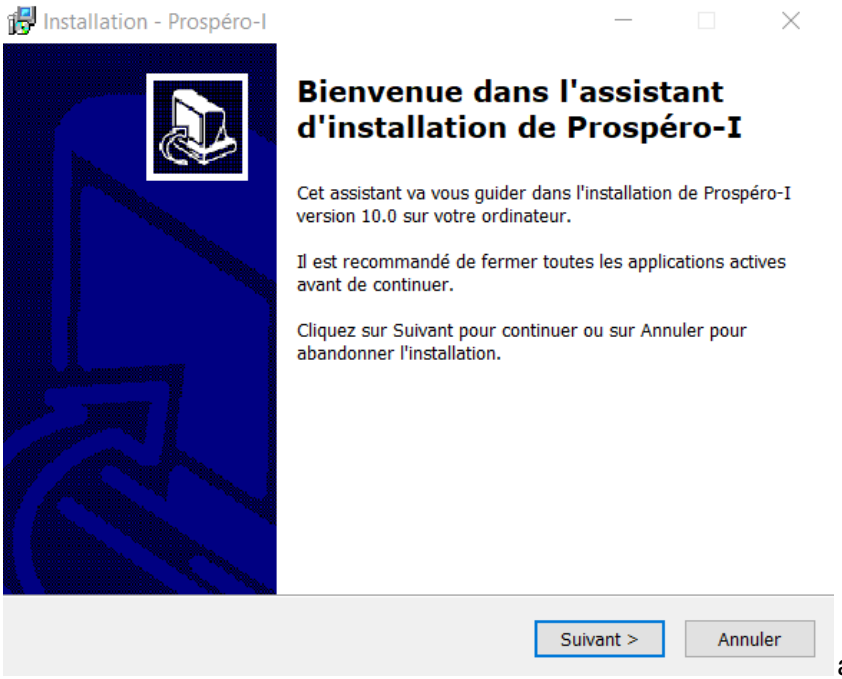
Po úspěšném dokončení registrace přijde na zadaný e-mail zpráva s odkazem ke stažení instalačního souboru. Tento soubor je možné získat také po přihlášení do uživatelského profilu přímo na stránkách Prospérologie, a to pod odkazem *Téléchargement* (stažení), verze *in Czech* (také na obrázku).

1.2 Instalace programu

Instalace Prospéra se příliš neliší od podobného postupu instalace u jiných programů. Níže nabízíme návod k instalaci krok za krokem.

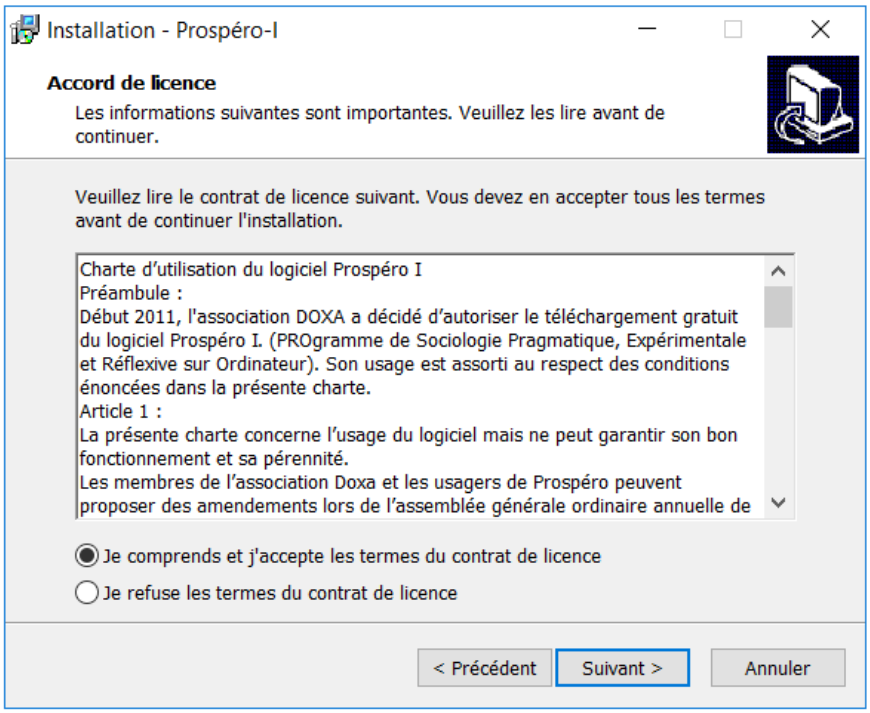
Při spuštění instalace (po stisknutí tlačítka Spustit) bude v některých případech třeba nejprve povolit aplikaci, aby prováděla na daném počítači změny (potvrdíme stisknutím tlačítka OK). Pro správnou funkčnost Prospéra také doporučujeme spuštění programu na Vašem uživatelském účtu "jako správce". Prospéro potřebuje příslušné povolení pouze k přepisování svých slovníků, tzn. nijak nezasahuje do jiných programů a aplikací.

Otevřením souboru setup–prospero–I–freeware.exe spustíte průvodce instalací programu Prospéro. Následně se otevře okno průvodce instalací, kde stiskneme tlačítko *Suivant*:

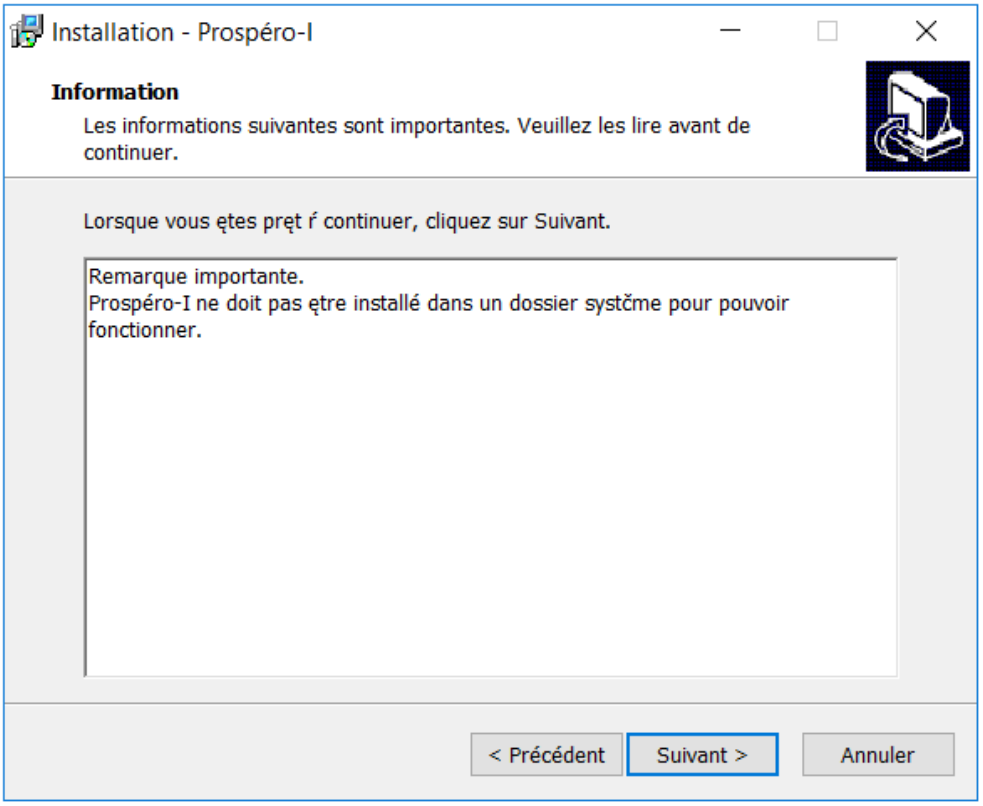


at

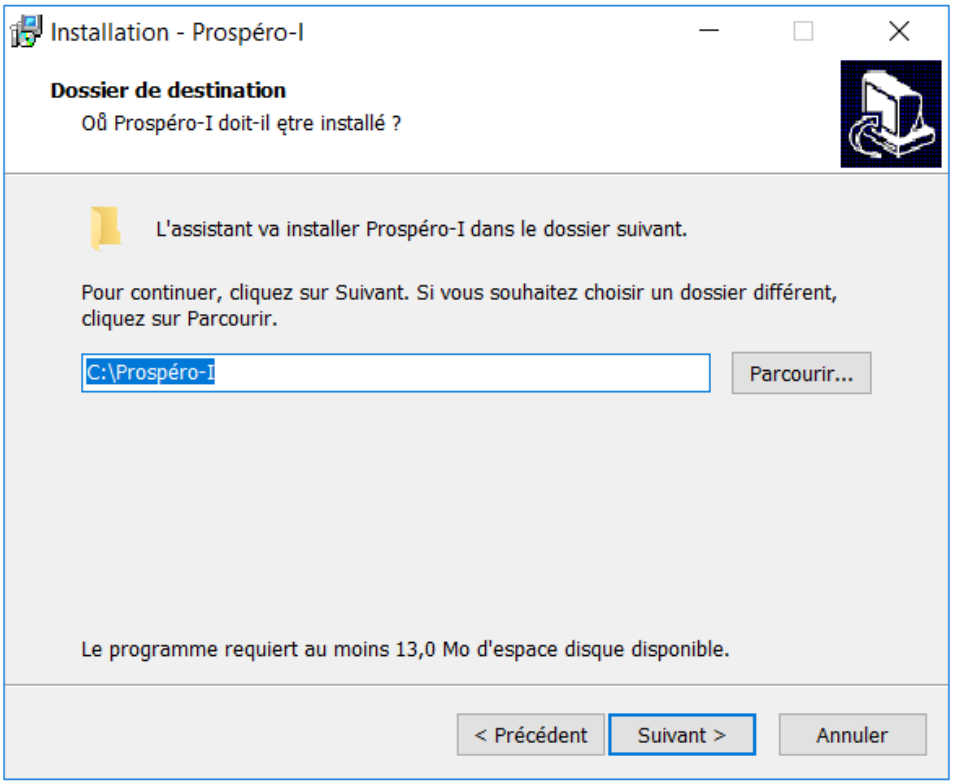
V další nabídce zaškrtneme pole *Je comprends et j'accepte les termes du contrat de licence* (Souhlasím s licenčními podmínkami) a pokračujeme stisknutím tlačítka *Suivant*:



Pokračujeme stisknutím tlačítka *Suivant*. Následuje upozornění k umístění instalace:



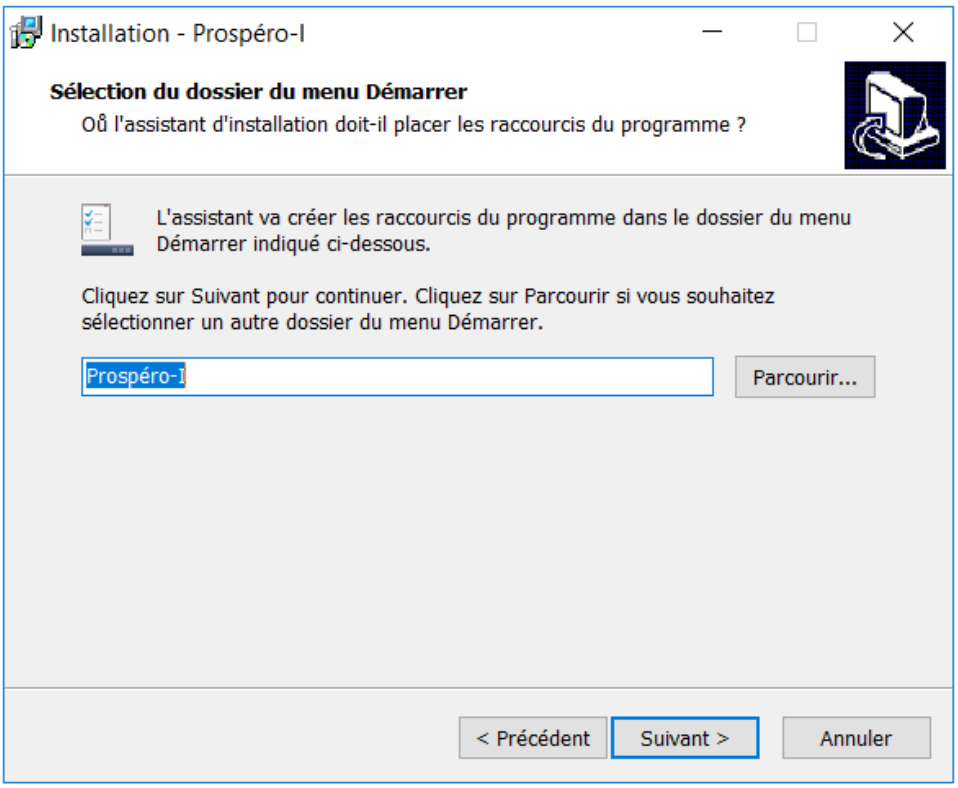
Pokračujeme opět stisknutím tlačítka *Suivant*. Umístění instalace je přednastavené do cílové adresy:



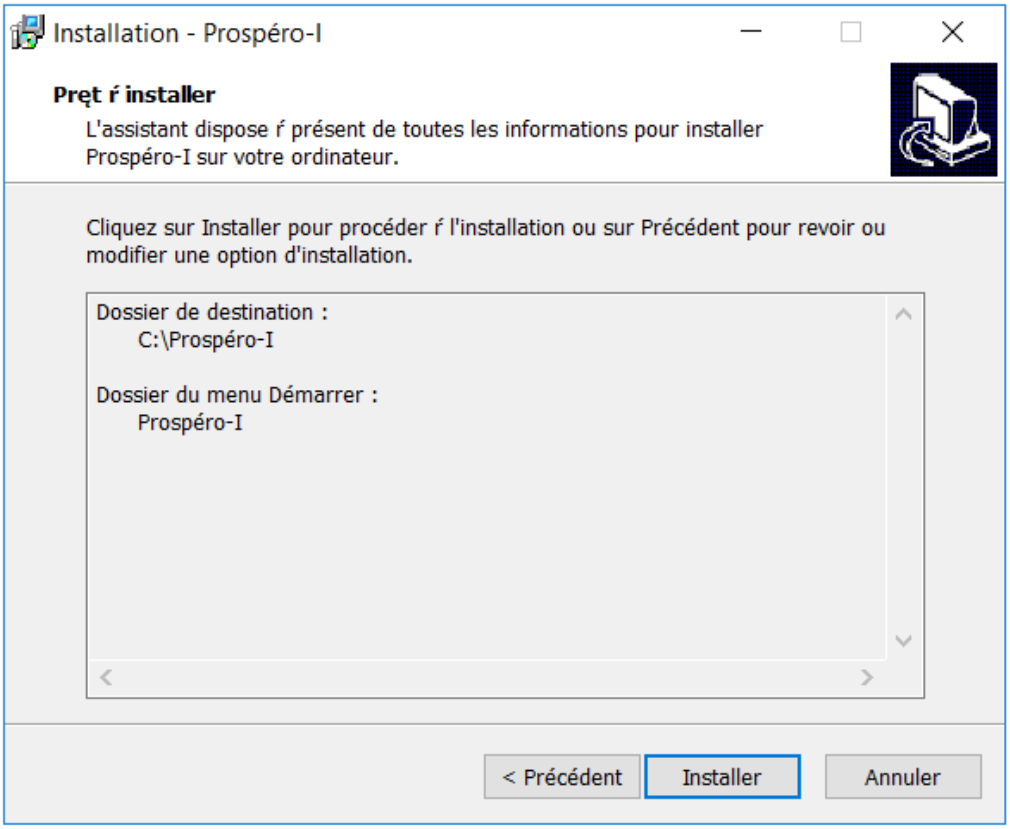
Správce instalace nabízí umístění instalace do nové složky C:\Prospéro-I. Toto umístění je doporučené – případnou změnu umístění provedete přes tlačítko *Parcourir…*)

Důležité upozornění: Aby program Prospéro pracoval správně, je v případě změny adresy instalace zapotřebí nastavit adresu umístění instalace na jinou, než je systémová složka operačního systému Windows.

Po vyplnění adresy instalace pokračujeme stisknutím tlačítka *Suivant*:



Nastavení se týká “zástupce složky” programu Prospéro (také přes nabídku Start). Zde pokračujeme stisknutím tlačítka *Suivant*:



Prospéro je nyní připraveno k instalaci.

Po stisknutí tlačítka *Installer* se program Prospéro nainstaluje do Vašeho počítače.

2. Příprava textů

Před samotným založením projektu a vkládáním textů do Prospéra nejprve představíme některé nutné (a doporučené) úpravy pro správnou funkčnost programu. Při přípravě článků (čištění, formátování) je třeba zohlednit dvě základní a několik dalších doporučených úprav korpusu.

Odstranění těchto nedostatků je třeba provést manuální úpravou textů; při větším počtu textů je pro usnadnění provedení úprav doporučeno nechat je v jednom souboru (nebo dočasně sloučit) a pak je rozdělit na jednotlivé texty pro vložení do Prospéra. K přípravě velkých korpusů z databáze *Newton Media* naleznete na stránkách Prospérologie odkaz s externím skriptem.

Mezi základní úpravy patří:

- Všechny články je třeba převést (uložit z databází) ve formátu .TXT, neboť Prospéro nerozezná jiné formáty textů (.doc, .docx, .odt, .pdf, .xls apod.).

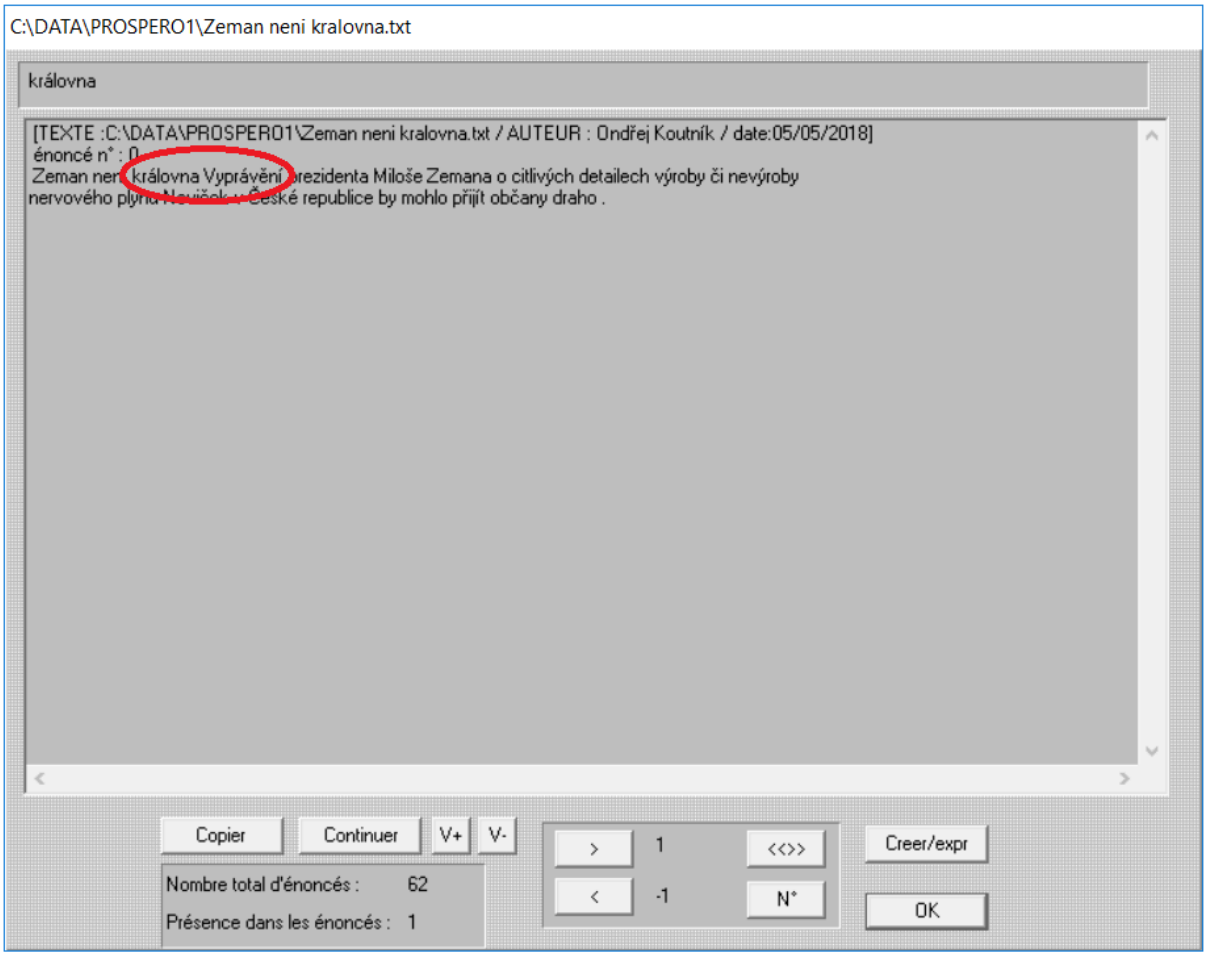
- Každý uložený text musí mít nastavené zalamování řádků. Toto nastavení můžeme zkontrolovat pro každý text otevřený v textovém editoru (např. v Poznámkovém bloku je tato volba na hlavním panelu pod záložkou Formát).
- Prospéro neumí pracovat s grafickým formátem obrázků – vložením do Poznámkového bloku dojde k jejich odstranění, zachovány jsou však jejich popisky.
- Prospéro rozpoznává českou diakritiku (háčky a čárky), tj. kódování jazykové sady specifikované normou ISO-8859-2. Nicméně stává se, že na pozici znaku uvnitř slova nebo mezery mezi slovy jsou přítomné jiné „neviditelné“ speciální znaky (a symboly). Jejich přítomnost nemusí být ve Wordpadu/Poznámkovém bloku přímo zobrazena, a tak je třeba v případě potíží obsahy ověřit (a „očistit“) pomocí některého textového editoru.

Pokud je nám lhostejná syntaktická úroveň analýzy a zajímá nás jen čistě sémantická stránka korpusu, můžeme rozhodnout, že další úprava textů je časově náročnější než její přidaná hodnota. V opačném případě je žádoucí provést sérii dalších úprav zvyšujících rozpoznávací schopnost i přesnost funkcí programu.

Mezi doporučené úpravy patří:

- Pokud ponecháme uvnitř textu autora, název média, počet stran, místo a datum atd., analýza bude těmito informacemi zkreslena (například pokud hledáme frekvence konkrétních slov a výrazů). Metaúdaje, tzn. nadpis, autor, datum apod., je třeba zachovat zvlášť (viz kapitola ke tvorbě metadat).
- Odstranění všech znaků a značek, které nejsou z hlediska analýzy relevantní. Např. znaky typu ***, --------, pokračování na str. X. Tyto a jiné autorské či mediální znaky a značky mohou rovněž zkreslovat výstupy (pokud nemají pro naši analýzu speciální význam).
- Jména osob ani nadpisy článků by neměly být psané celé velkými písmeny. Prospéro rozlišuje mezi počátečními velkými a malými písmeny, nerozlišuje však diakritiku u počátečních písmen ve vlastních jménech, např. Čestmír. Ponecháme-li kapitalizovaná první písmena křestních jmen a příjmení, dokáže Prospéro rozpoznat, že se jedná o jména (aktérů) (viz oddíl Jména a příjmení). Je samozřejmě možné, že budeme chtít ponechat velká písmena tam, kde cítíme, že kapitalizace něco naznačuje o intencionalitě autora.
- V případě rozhovorů stylem otázka – odpověď nabízí program Prospéro možnost přiřadit výroky k jednotlivým mluvčím, ale jen pokud používáme zvláštní notaci (viz níže).

- Prospéro nerozpozná citace, pokud neupravíme formát uvozovek na anglický. Takto upravené uvozovky: “[CITACE]” obsahují obě značky v horní pozici. V poznámkovém bloku (Notepadu) se tyto uvozovky zpravidla zobrazují rovné.
- Interpunkční znaménka typu čárka/tečka (konce citace) umísťujeme po uvozovkách.
- Prospéro nerozpozná konec věty (včetně citací), pokud končí třemi tečkami. Řešením je vložení mezery mezi slovo a tři tečky. V opačném případě Prospéro čte tři tečky jako součást slova. Mezera musí být umístěna i před další přidanou tečkou, která označuje konec věty. Například ***Povídejte…*** se změní na ***Povídejte … .***
- Je také třeba umístit tečku po nadpisech a podnadpisech (a také mezi řádky odrážkového seznamu), jinak je Prospéro spojí s následující větou. Například pokud úvodní nadpis článku „Zeman není královna“ neobsahuje tečku, pak Prospéro tento nadpis neoddělí od následující věty „Vyprávění prezidenta...“ (viz obrázek níže).



Pro pochopení úprav textů, které předchází založení projektu, je nejlepší postupovat na základě krátkých cvičných příkladů. K tomuto účelu byly vybrány dva texty, jeden

neupravený a jeden upravený, a také jeden příklad upraveného textu, který obsahuje rozhovor.

Příklad neupraveného textu:

Metadata: *Zeman není královna,* Lukáš Koutník, Lidové noviny, 5. 5. 2018

Zeman není královna
Prezident republiky má dle ústavy trestní imunitu a není z výkonu své funkce odpovědný. „Britské královně nemáte problém svěřit tajemství, protože nepředpokládáte, že ho bude vykládat po náměstích a do televize," poznamenává ústavní právník Jan Kysela. „V zásadě se má za to, že odpovědně by postupovala každá hlava státu. Pokud se čirou náhodou stane, že hlava státu z takového očekávání vybočí, zdá se mi, že nejde o vadu systému, ale exces, se kterým nejde nic příliš dělat," komentuje případ Kysela.
Dodává, že možností, jak do budoucna zabránit podobným problémům s vyzrazením tajných informací, je úprava zákonů. „Z ústavy neplyne, že se máte dostávat ke zprávám tajných služeb. To vyplývá až z různých zákonů. Můžete si říci, že hrozí-li vám riziko, že se stanete nedůvěryhodnými pro partnery v EU a NATO, nebudete určité informace prezidentovi poskytovat. Je to sice smutné, není to systémové, ale je to možná lepší než sedět se založenýma rukama," míní Kysela.

Příklad upraveného textu:

(provedené úpravy jsou označeny červenou barvou)

Metadata: *Novičok v Česku*, Eva Pospíšilová, Alžběta Šimková, MF Dnes 6. 5. 2018 (upraveno)

** Ruská federace Zemanův výrok ihned přivítala **.
V České republice slova prezidenta Miloše Zemana vyvolala pozdvižení. Kdo je naopak přivítal, bylo Rusko. Podle mluvčí tamního ministerstva zahraničí Mariji Zacharovové totiž dokazují, že Británie v kauze Skripal lže. To by podle ní měla vysvětlit, proč tajila informace o otravné látce novičok v Česku.
V České republice však Zeman pochopení nenašel. "Miloše Zemana můžeme klidně považovat za aktivního agenta Kremlu. Nejenže svým prohlášením odporuje českému ministrovi zahraničí, navíc tím živí ruská propagandistická média", uvedl předseda hnutí STAN Petr Gazdík. Šéf poslaneckého klubu TOP 09 Miroslav Kalousek však zdatně sekundoval. "Miloš Zeman je na roztrhání. Čínští investoři ho mají na dobré vazby na Kreml. Rusové ho mají na svou lživou propagandu. My ho máme na permanentní ostudu", uvedl na svém twitterovém účtu. Předseda ČSSD Jan Hamáček pak uvedl, že za klíčové pokládá sdělení, že v tomto případě se jednalo o látku podobnou (nikoliv identickou) látce, která byla použita k atentátu v Salisbury. "Tudíž ČR nemůže být jakkoliv zmiňována jako země, odkud by použitá chemikálie měla pocházet", napsal MF DNES Hamáček. Podle premiéra v demisi Andreje Babiše je překvapující, že se tajné služby rozcházejí ve vyjádření o přítomnosti nervového jedu novičok v Česku. "Takže se jich musíme zeptat, jak to vlastně myslely", řekl Babiš. Prezidenta se včera nicméně zastal na Twitteru jeho mluvčí Jiří Ovčáček: "Jak za časů normalizační KSČ. Pravda je nežádoucí. Důležitý je správný ideový postoj".

III. Práce s projektem

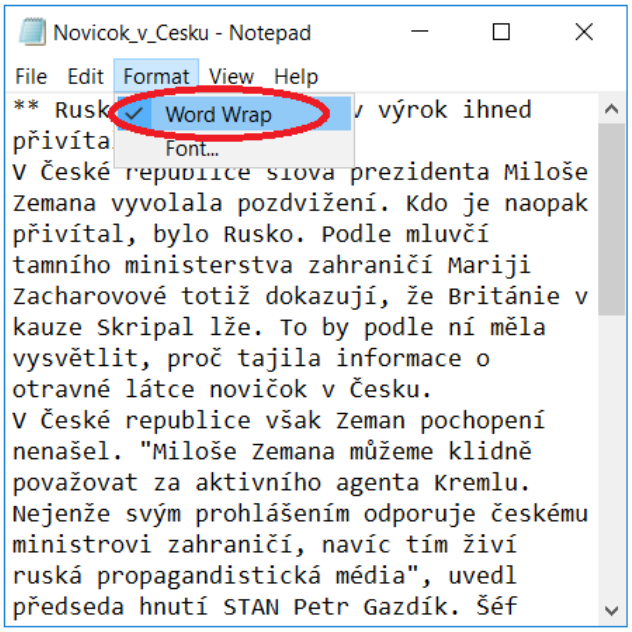
1. Vytvoření projektu

a. Příprava souboru .TXT

Prvním krokem založení projektu je příprava textů ve formátu .TXT. Pro účely demonstrace využijeme dvou předchozích cvičných textů („Zeman není královna“ a „Novičok v Česku“), které vložíme do prázdného souboru textového editoru Poznámkový blok (Notepad). Editor je obvykle součástí výbavy operačního systému Windows, a prázdný soubor tak lze vytvořit například pravým tlačítkem myši v průzkumníku souborů.

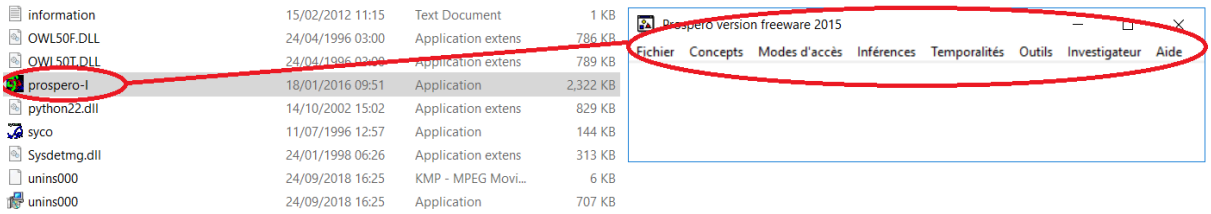
K založení textů (korpusu) lze využít několik možností. V libovolném textovém editoru můžeme uložit prázdný soubor ve formátu .TXT. Jiný způsob je otevřít již některý existující soubor .TXT a ten přepisovat a ukládat pod novými jmény jednotlivých textů.

Při vložení/kopírování textu do editoru je třeba zkontrolovat zalamování řádků (obrázek níže). Toto nastavení ověříme v záložce „Formát – Zalamování řádků“ (angl. „Format – Word Wrap“). Pokud zvolíme postup přípravy textů průběžným ukládáním jednoho souboru pod novými jmény, toto nastavení by mělo zůstat zachováno pro každý další uložený text.

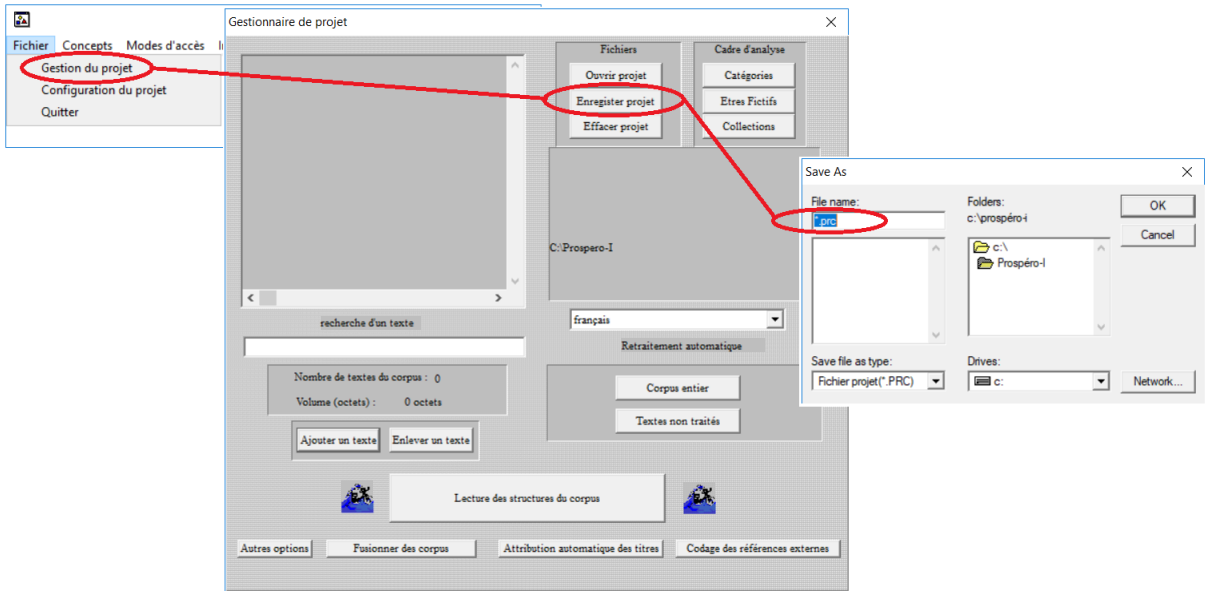


b. Založení projektu, vkládání textů

Založení projektu je jedna z prvních operací, kterou provádíme přes uživatelské rozhraní (existuje ještě jiná alternativa, ale k tomu až později). Nejprve spustíme program Prospéro pomocí spouštěcího souboru *prospero-I.exe*, který se nachází v adresáři instalace. Po spuštění se nám zobrazí následující okno – základní rámec uživatelského rozhraní.



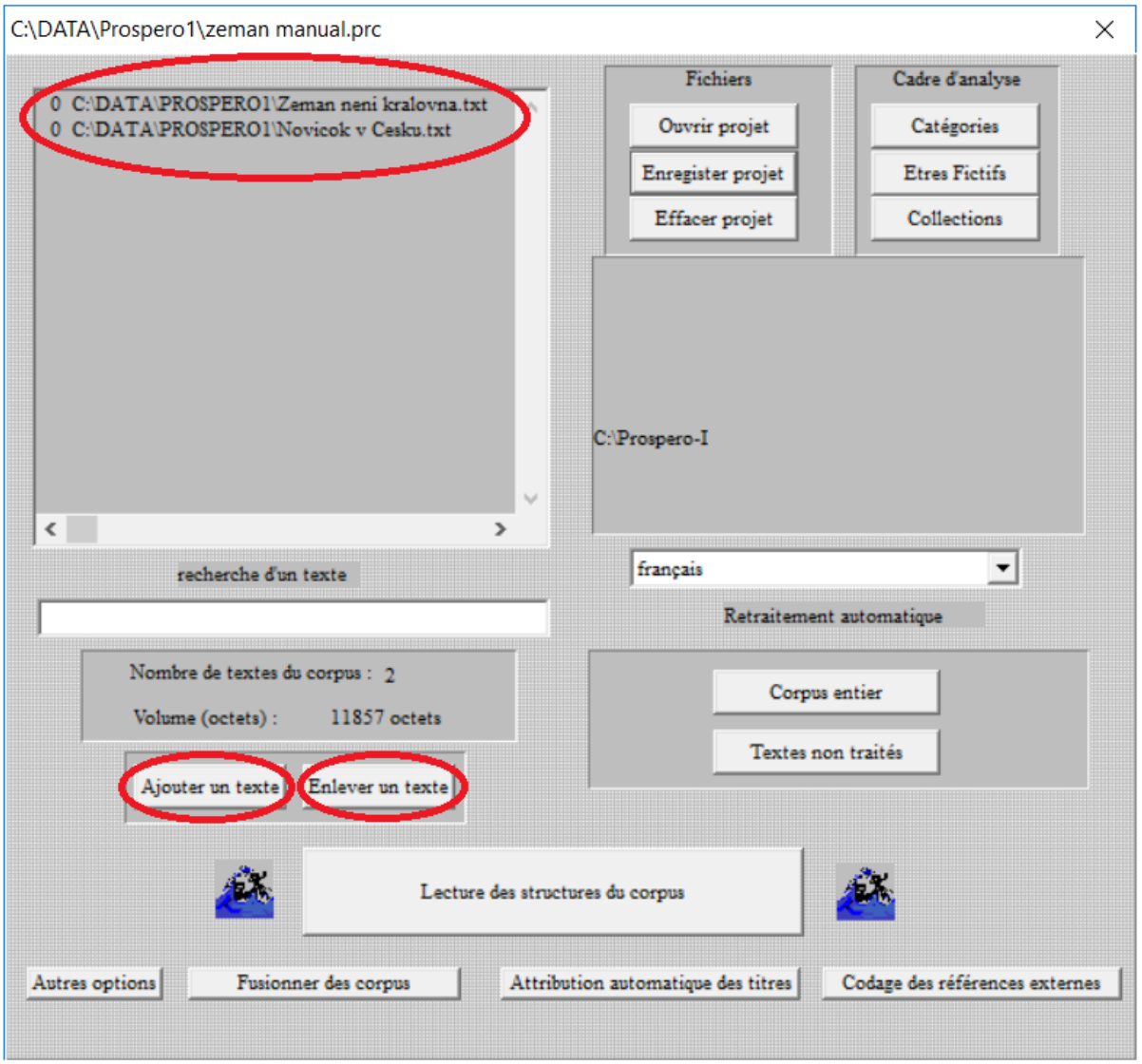
Následuje otevření okna správy projektu (*Gestion du projet*) v menu Soubor (*Fichier*). Pod tlačítkem uložení projektu (*Enregistrer projet*) se otevře nové dialogové okno, ve kterém projekt zakládáme vytvořením příslušného souboru .PRC



Pro názornost si vytvoříme projekt pod názvem *zeman manual.PRC* (název vyplníme do volného pole s příponou *.prc) Při dalším spuštění Prospéra máme možnost tento projekt otevřít (*Ouvrir*), nebo smazat (*Effacer*) jeho veškerý obsah. Tento obsah, jednotlivé texty korpusu, je ale třeba nejprve doplnit/připojit k projektu.

Texty vkládáme jednotlivě přes uživatelské rozhraní pod záložkou správy projektu (*Gestion du projet*) v menu Soubor (*Fichier*), dle obrázku níže. U větších korpusů doporučujeme přetažení označené skupiny textů z okna prohlížeče do příslušného místa v okně správy projektu. Tím se ušetří čas oproti jednotlivému vkládání textů.

V záložce správy projektu máme možnost kliknout na kterýkoli text, který chceme odstranit/smazat (*Enlever un texte*), nebo korpus doplňovat vkládáním dalších textů *(Ajouter un texte*).



Projekt s nově vloženými texty je vždy nutné průběžně **ukládat** tlačítkem *Enregistrer projet* v rozhraní správy projektu (*Gestion du projet*) .

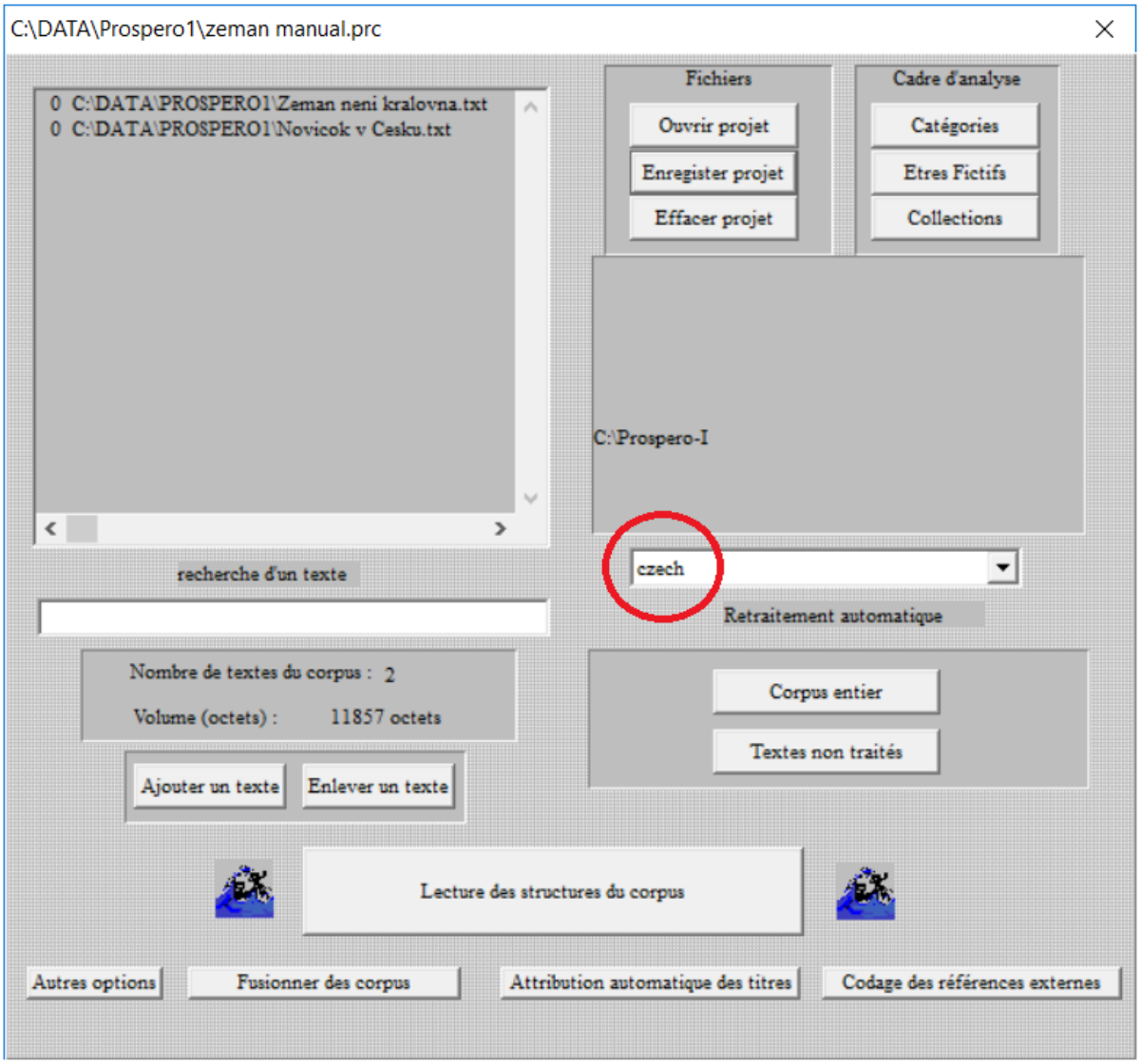
c. Vkládání jazykových slovníků

Soubor .PRC obsahuje nejen seznam textů celého korpusu s jejich adresami umístění, ale také adresu umístění jazykových a pojmových slovníků.

Po úspěšné instalaci české verze Prospéra byl k programu dodán veškerý příslušný **jazykový obsah**, tj. klíč, pomocí kterého bude schopen číst české texty, vyhledávat či srovnávat jejich obsahy. Konkrétně se zde jedná o slovník „dictio.cfg“.

Slovník „dictio.cfg“ je jednou ze základních jednotek programu Prospéro. Do instalační složky programu je umístěn automaticky, tzn. již při samotné instalaci české verze programu:

V případě instalace jiné jazykové verze (francouzské, anglické) je nutné soubor fyzicky přepsat českou verzí slovníku. V rámci uživatelského rozhraní je pak třeba u každého nového projektu (v našem případě „zeman manual.PRC“) změnit **nastavení dictio.cfg** ve správě projektu (*Gestion du projet*). Doporučujeme toto nastavení následně uložit (*Enregistrer projet*).



Další jazykové slovníky

Zatímco *dictio.cfg* slouží k identifikaci slovních druhů na základě pravidelností jejich typických koncovek (např. adjektiva s koncovkami *–ý, -á, -é*), další slovníky tuto klasifikaci slov zpřesňují zohledněním nepravidelných nebo nejednoznačných koncovek (tzv. homonymní výrazy) či frazémů (viz oddíl Vytváření frazémů). Konkrétně se jedná o následující slovníky:

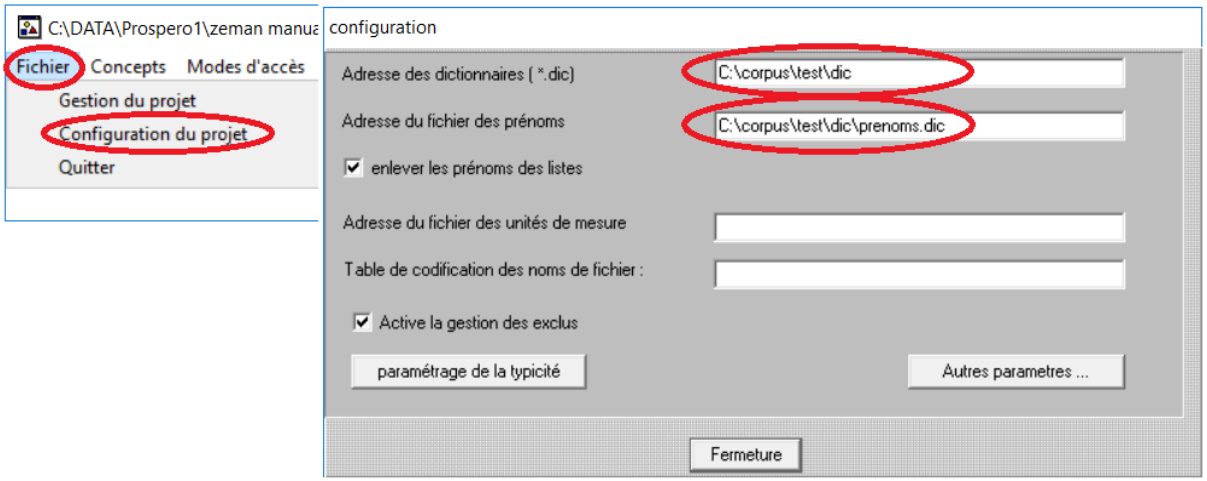
- cz_epreu.DIC obsahuje slovesa,
- cz_etre.DIC obsahuje podstatná jména,
- cz_nomb.DIC obsahuje číslovky,
- cz_outi.DIC obsahuje spojky a zájmena,
- cz_quali.DIC obsahuje přídavná jména,
- cz_marqu.DIC obsahuje příslovce, konektory, argumentační částice,

- prenoms.DIC obsahuje křestní jména osob.

Česká instalace Prospéra opět nabízí příslušné české slovníky, umístěné ve složce:

C:\corpus\test\dic

Nastavení jazykových slovníků se vždy vztahuje ke konkrétnímu projektu (tato informace je uchována, jak již bylo uvedeno, v souboru .PRC). Proto je třeba u každého nového projektu propojit slovníky, resp. nastavit cestu ke slovníkům. V rámci uživatelského rozhraní programu tuto cestu nastavujeme přes záložku *Configuration du projet* v menu *Fichier*:



a) řádek *Adresse des dictionnaires* (společná cesta k jazykovým slovníkům)

C:\corpus\test\dic

b) řádek *Adresse du fichier des prénoms* (cesta k seznamu (křestních) jmen - souboru prenoms.DIC - uloženému ve složce *dic*:

C:\corpus\test\dic\prenoms.dic

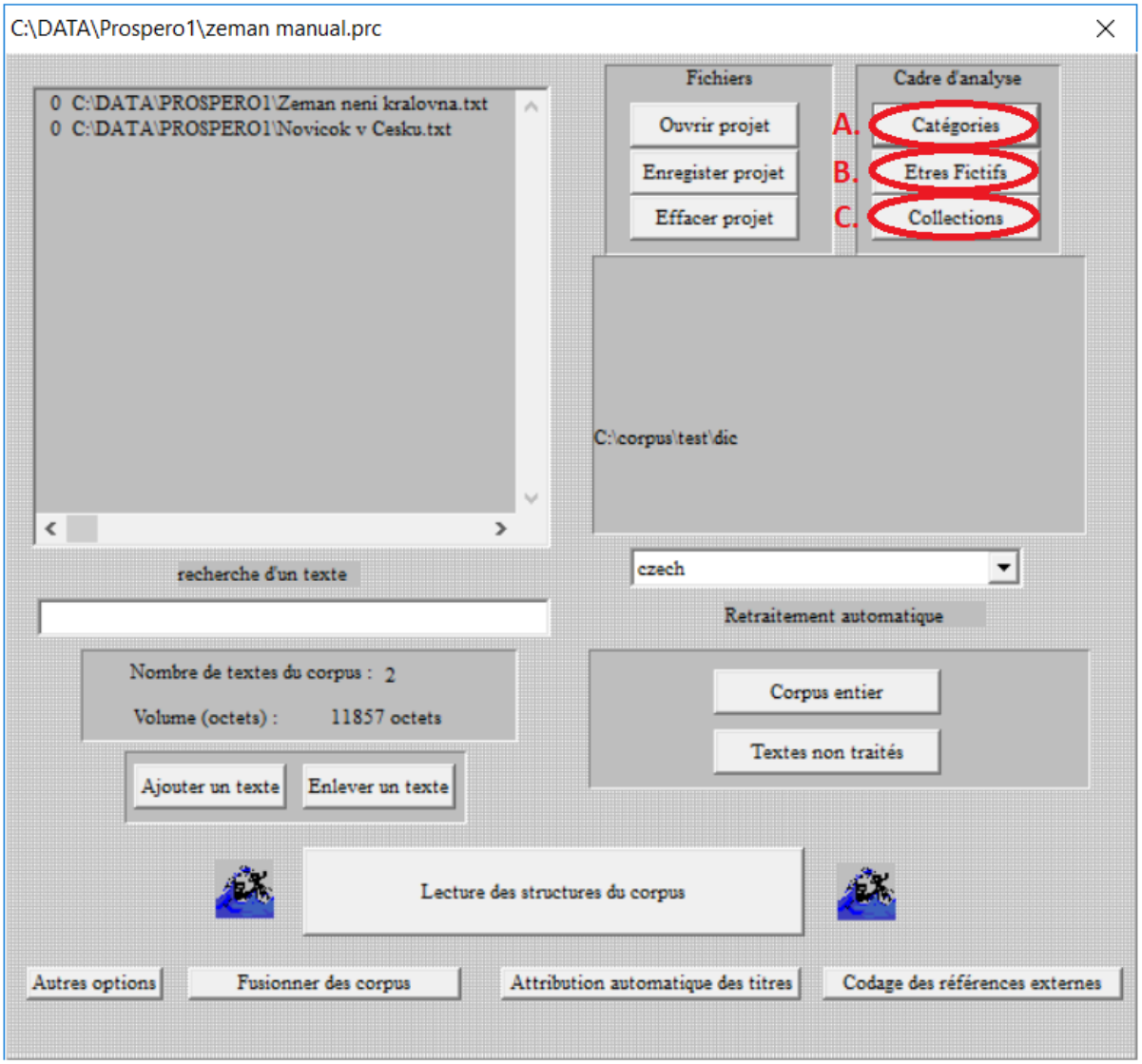
Ostatní pole můžete ponechat volná. Ještě předtím, než okno zavřeme (*Fermeture*), doporučujeme zaškrtnout *Enlever les prénoms des listes* a také *Active la gestion des exclus*. Výsledkem tohoto dílčího nastavení křestní jména osob zmizí ze seznamu Entités, aby nebyli aktéři do analýzy zahrnuti hned dvakrát.

d.Vkládání pojmových slovníků

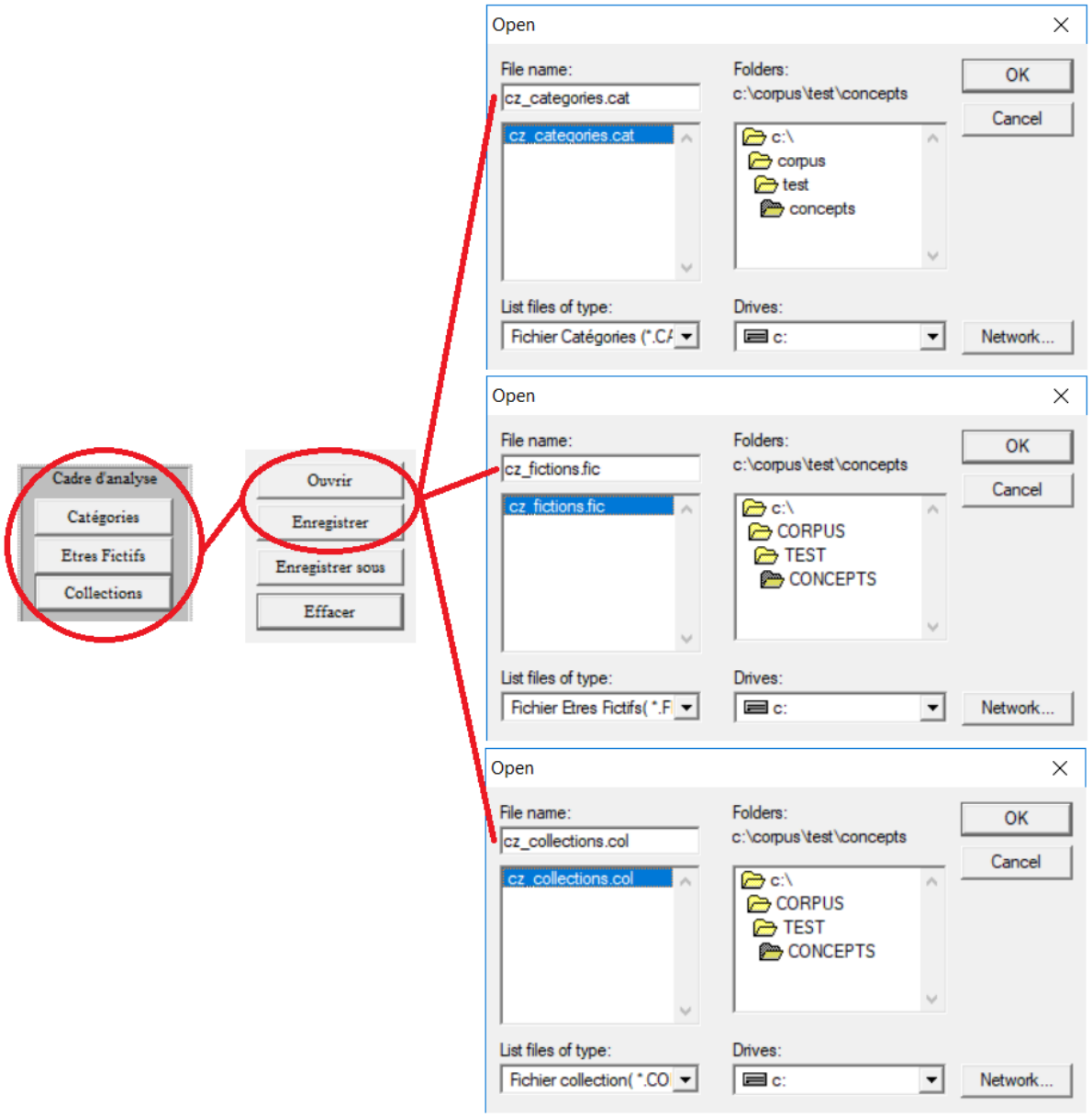
Posledním krokem před samotnou analýzou je propojení vznikajícího projektu s pojmovými slovníky, které zastupují konceptuální úrovně analýzy:

- kategorie (soubor s příponou .CAT),
- velké postavy (soubor s příponou .FIC),
- sbírky (soubor s příponou .COL).

V základním rozhraní správy projektu (*Gestion du projet*) se zaměříme na nastavení tohoto rámce analýzy (*Cadre d'analyse*), kterému je vyhrazena pravá horní část rozhraní. Nejprve (A) nastavíme cestu ke slovníku s kategoriemi (tlačítko *Catégories*). Následně (B) doplníme slovník s velkými postavami (tlačítko *Etres Fictifs*). Na závěr (C) doplníme slovník se sbírkami (tlačítko *Collections*).



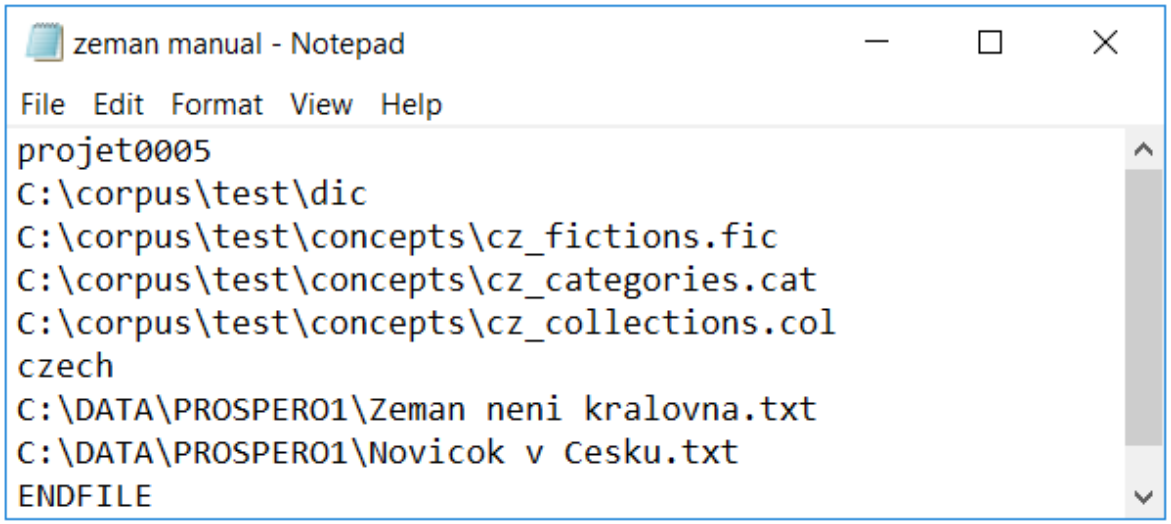
Propojení se vždy vztahují ke konkrétnímu projektu. Zároveň platí, že nejsou nutně výsledkem kódování korpusu. Opět využijeme vzorové příklady.



Přes záložku rámce analýzy (*Cadre d'analyse*) a otevření správce konceptů (*Catégories*, *Etres Fictifs*, *Collections*) se postupně dostaneme k dialogovému oknu, kde provedeme nastavení cesty k příslušnému souboru. Tato cesta vede do složky Dokumenty, v níž nalezneme složku s uloženými pojmovými slovníky (tato složka vznikla během instalace příslušné jazykové verze programu Prospéro). Je důležité vždy potvrdit nastavení uložením (*Enregistrer*).

Přímá editace souboru .PRC mimo uživatelské rozhraní: pokud jsme uložili vytvořený projekt, vložili texty a nastavili cestu ke slovníkům, pak již má Prospéro všechna potřebná data pro analýzu. Tento soubor nově vytvořeného projektu (soubor uložený s příponou .PRC) máme možnost otevřít (a také editovat) mimo uživatelské rozhraní, tzn.

přímo z umístění na pevném disku. Soubor .PRC se otevírá v Poznámkovém bloku a jeho zdrojový kód je v následujícím formátu.



Tato informace je užitečná, pokud chceme doplňovat seznamy textů z různých umístění anebo naopak mazat rozsáhlé části korpusu. Pokud například přesuneme texty, nebo dokonce některé slovníky, mezi složkami počítače, musíme zároveň upravit všechny jejich adresy v projektu. Pracovat s obsahem souboru .PRC přímo zde může být rychlejší variantou (např. přes funkci Najdi a nahrad). Struktura souboru je následující:

- První (*projet0005*) a poslední (*ENDFILE*) řádek značí začátek a konec projektu – tato pole musí obsahovat každý .PRC soubor v nezměněné podobě.
- Druhý řádek označuje cestu k jazykovým slovníkům (tj. *entité*, *qualité*, *épreuve*).
- Následují adresy tří pojmových slovníků, které jsme připojili ke korpusu (v tomto pořadí: *Etres fictifs*, *Catégories, Collections*).
- Další řádek označuje za jazyk korpusu (*dictio.cfg*) češtinu.
- A konečně jsou zde za sebou řazena data, tj. naše jednotlivé texty – vždy jeden text na řádek – včetně celé adresy umístění na disku.

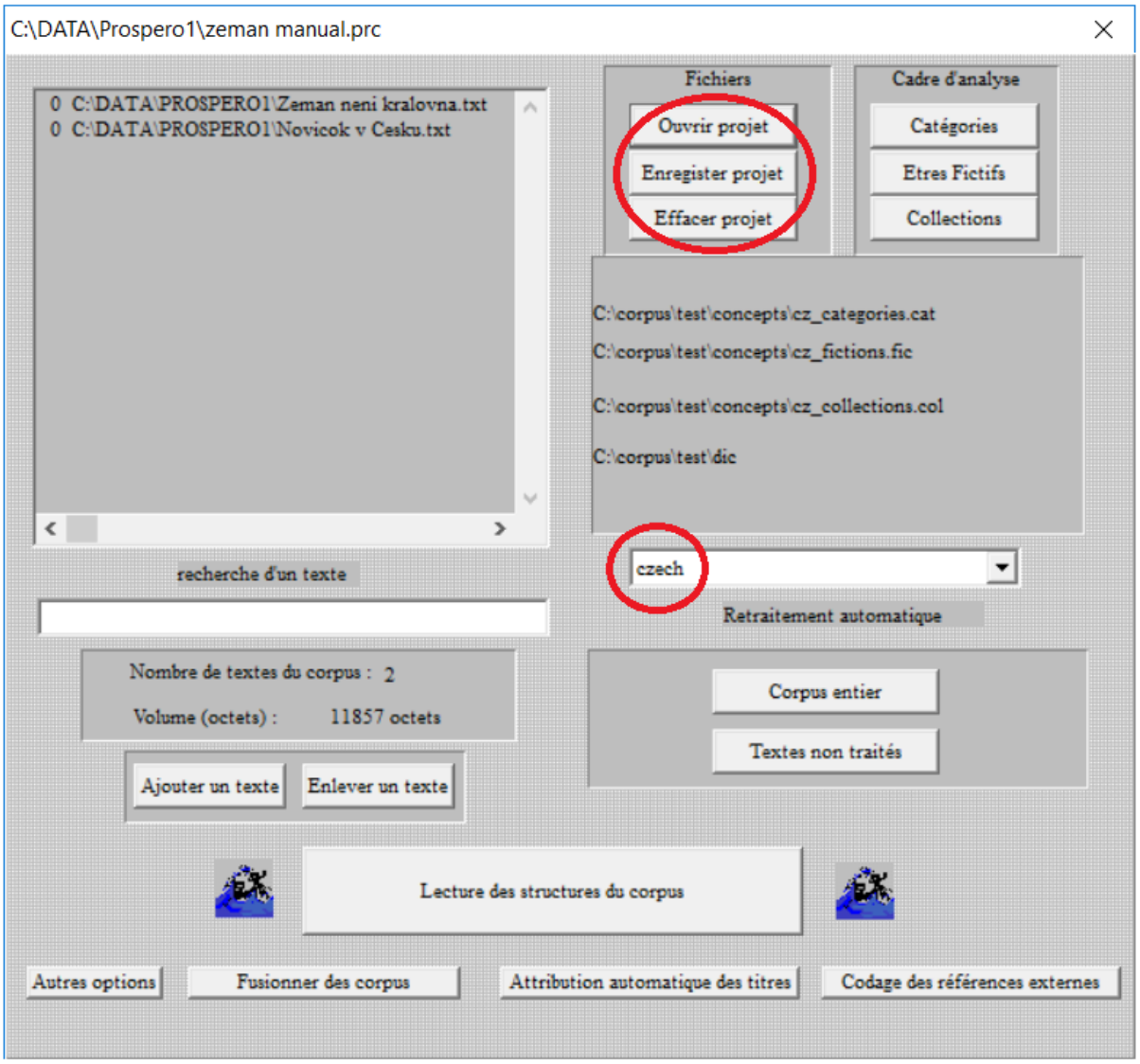
Pozn.: Nyní lépe chápeme, proč a kdy je třeba projekty průběžně ukládat: vždy, když přidáme, odebereme nebo přemístíme jeden či více z uvedených souborů.

e. Rekapitulace, první spuštění

Pokud jsme již jednou založili projekt (a také jej uložili přes *Enregistrer projet*), pak tento projekt můžeme kdykoliv otevřít/smazat příslušným tlačítkem (*Ouvrir projet*/*Effacer projet*) na rozhraní správy projektu (*Gestion du projet*).

V levé části rozhraní máme seznam všech textů projektu a také informaci o jejich umístění. Vpravo uprostřed jsou zobrazeny cesty k jazykovým a pojmovým slovníkům (okno se nepřizpůsobuje délce adresy, a tak ji často nevidíme celou). Pod touto částí rozhraní

máme možnost zkontrolovat (a případně změnit) nastavený jazyk analýzy, v našem případě češtinu (viz také obrázek níže).



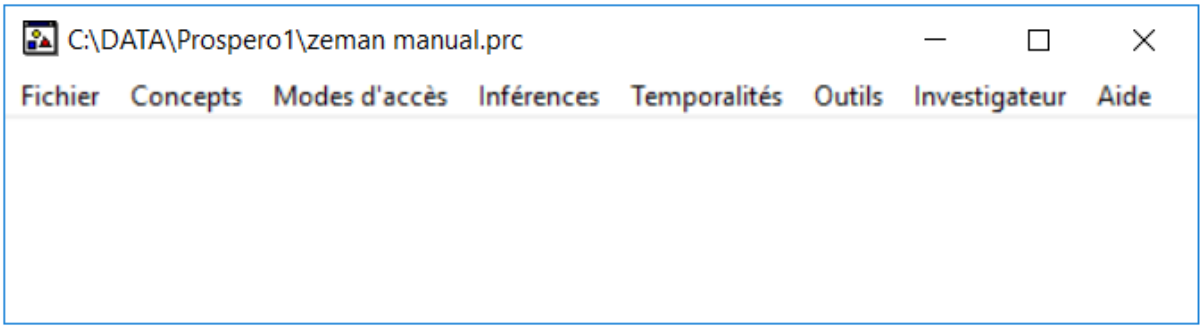
Načtení struktury korpusu: všechna předchozí nastavení (jazyka, slovníků, korpusu) zůstávají zachována. Po prvním spuštění i při většině prováděných úprav (korpusu/analytických kategorií) je však třeba „načíst“ veškerý obsah do aktivní paměti programu. Například pokud jsme se rozhodli doplnit nové texty, nebo provádíme-li další změny (zejména typizace slov, vytváření konceptů, zařazení slov ke konceptům), je třeba:

1. potvrdit Prospéru označení celého korpusu (tlačítkem *Corpus entier*), nebo doplnění chybějících článků (tlačítkem *Textes non traités*),
2. provést (na)čtení úprav (tlačítkem *Lecture des structures du corpus*). To může trvat několik sekund (občas se stane, že Prospéro „spadne“). Po zaznění signálu Windows (cinknutí) je program připraven k analýze.

Číslice 1 a 0 vedle článků indikují, je-li článek načten v mezipaměti za účelem analýzy. Při čtení velkých korpusů se stane, že Prospéro články vynechá. V takovém případě je třeba nenačtené články indikované číslicí 0 opětovně načíst přes tlačítko *Textes non traités*.

Základní rozhraní/nabídka: na obrázku níže se nachází základní rozhraní Prospéra – vstupní brána k analýze. V záhlaví této hlavní nabídky máme informaci o současném projektu, v našem případě „zeman manual.PRC“. Hlavní záložky označují přístup ke slovníkům, analýze na lingvistické úrovni a socio-diskurzivní analýze z hlediska různých inferencí, temporalit a dalších nástrojů.

Nicméně Prospéro zde má natolik provázanou strukturu (cesty k různým funkcím se překrývají), že bude nejlepší postupovat k těmto funkcím z hlediska „úrovně“ analýzy. Předmětem našeho výkladu tedy nebude, „co které tlačítko dělá“ (hlavně pro nefrankofonního uživatele to může být ze začátku obtížné zjistit), ale jak se dostat na tyto jednotlivé úrovně analýzy.



Pozn. některá okna se zavírají tlačítkem OK, jiná křížkem nebo tlačítkem *Fermer*, které jsou umístěné nahoře i dole. Data se ovšem zavřením neztratí, dokud nezavřeme celý projekt nebo program.

2. Analýza na lingvistické úrovni

Výsledkem vložení korpusu do Prospéra a načtení jeho struktury je jazyková indexace a pojmové třídění. Tento a následující oddíl (Socio-diskurzivní analýza) slouží k obeznámení s fungováním jazykových a pojmových slovníků, na jejichž základě se tyto klasifikace dějí, a s možnostmi jejich úpravy.

Jazykové slovníky obsahují slova v základních tvarech, ale také celá jejich paradigmata, samostatně uvedené koncovky či přípony tak, aby program rozlišil co nejvíce jazykových

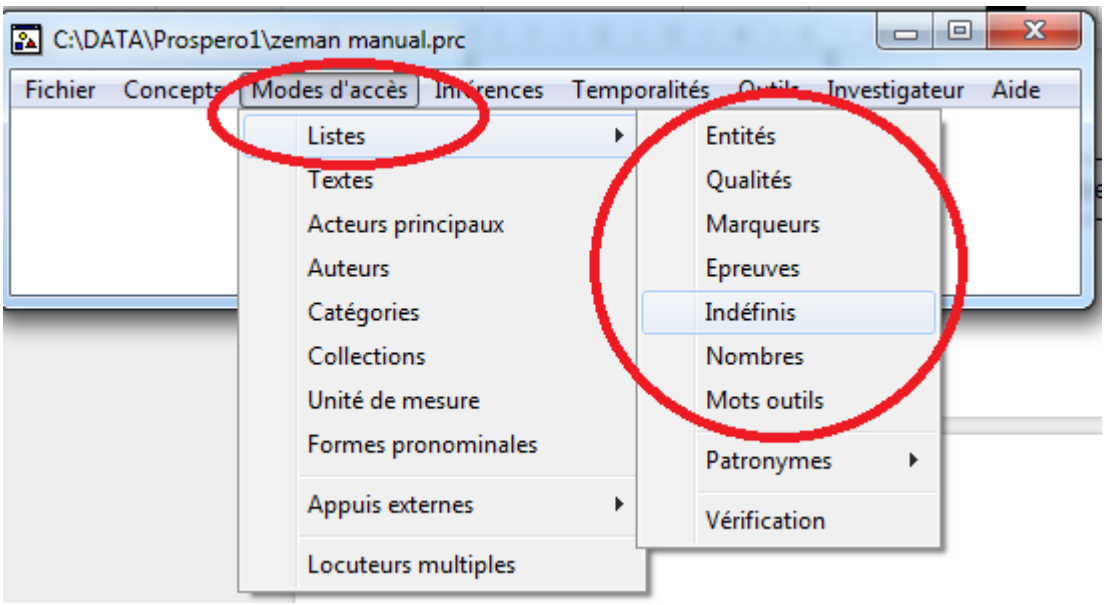
jednotek (případy, kdy se vyskytnou např. homonymní výrazy, je třeba řešit individuálně dle potřeb projektu, blíže viz kapitola Slovní kategorizace a struktura slovníků).
Následující tabulka představuje způsob kategorizování jednotlivých lexémů v Prospéru a termíny, kterými jsou zde označovány.

Druh slova nebo frazému	Nejbližší jazyková kategorie	Analytické zaměření
entité	podstatné jméno	témata, aktéři, aktanti, jevy, substance, entity, abstrakce, stavy – věcný „obsah“ diskurzu
qualité	přídavné jméno	příznaky, vlastnosti, kvalifikace, kvality
épreuve	sloveso	činnosti, děje, stavy, transformace, hodnocení – výrazy etablující vztahy mezi *entités*
marqueur	příslovce, příslovečné určení	bližší okolnosti dějů, vlastností; znaky modalizace, argumentační konektory (místa, času, způsobu atd.); mají zpravidla širší platnost než slovní druh příslovce
mot-outil	funkční slova: spojka, zájmeno apod.	spojení slov a větných členů vyjadřujících jejich vztahy, rozčlenění vět; zastoupení podstatných či přídavných jmen, odkaz na ně, jejich bližší určení
nombre	číslovka	označení počtu, pořadí, množství, měření času nebo velikosti / fyzikální jednotky (času, objemu, délky apod.)
indéfini	dosud neznámý řetězec písmen	dosud neklasifikovaný výraz / neklasifikovaná množina slov / novotvar / blíže neurčená slova (nevyskytující se v žádné jiné kategorii)

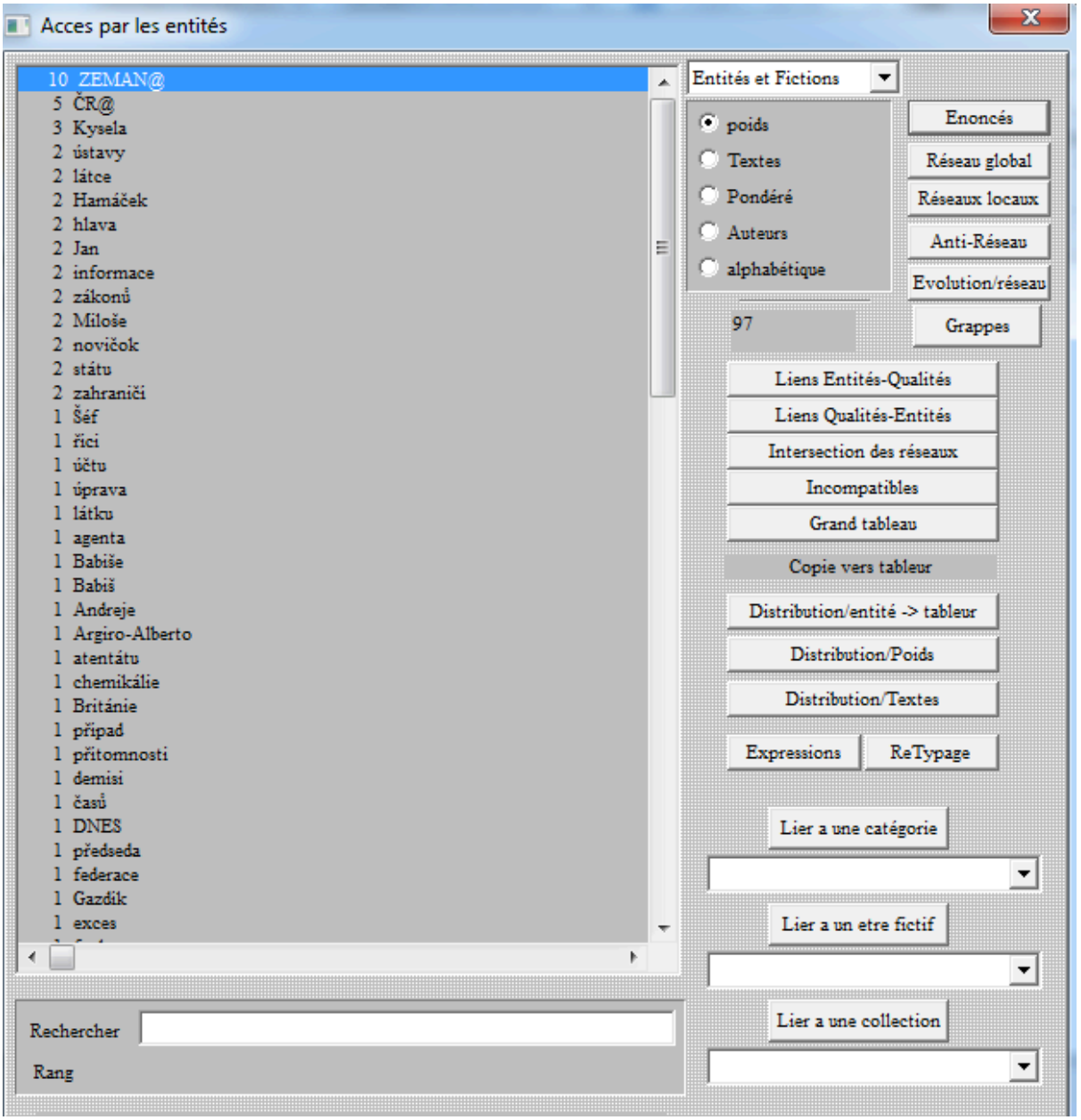
a. Kategorizace slov a struktura jazykových slovníků

(Menu *mode d'accès – listes*)

Do rozhraní pro seznamy slov, s nímž budeme pracovat v této kapitole, přistupujeme z hlavního menu Prospéra takto:



Pozn.: Pro *entités* vypadá rozhraní jinak než pro ostatní jednotky, je výrazně rozmanitější a nabízí více funkcí. Pouze podstatná jména lze například řadit i abecedně (*alphabétique;* vpravo nahoře):



Mezi *entités* řadí Prospéro i zkratky (v našem projektu např. STAN, EU, DNES, NATO). Odlišným jevem jsou velkými písmeny a následným znakem @ zapsané tzv. *Etres fictifs* neboli velké postavy korpusu (ZEMAN@, ČR@, OVČÁČEK@ atd., blíže viz příslušná kapitola).

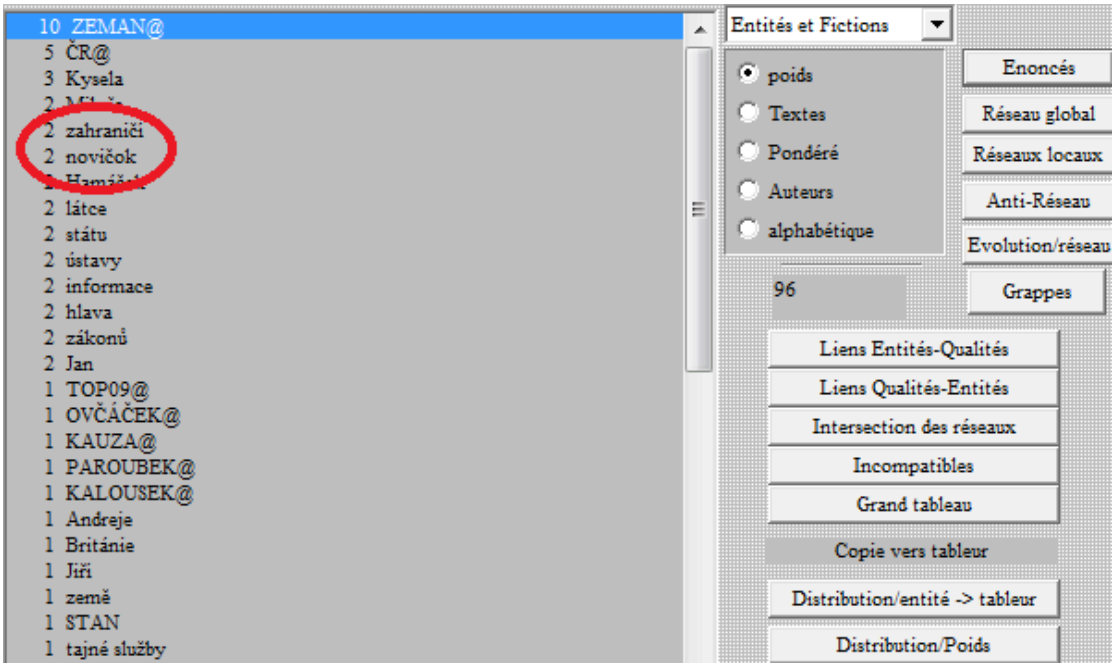
Ukázka rozhraní pro *marqueurs*:

Všimněme si, že se u *marqueurs* zde dle příkladů jedná o významově širší pojetí než pouhý slovní druh příslovce.

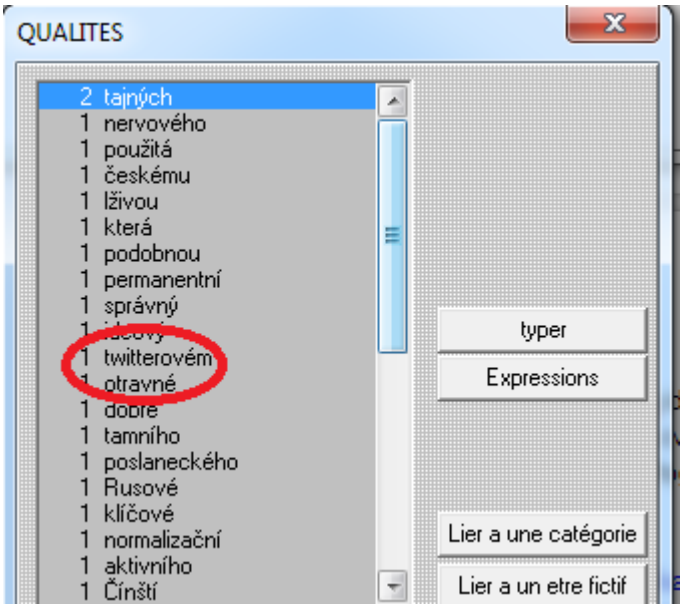
i. Typy klasifikace slov

Jak Prospéro čte vložené texty? Prospéro dokáže správně identifikovat slova třemi způsoby:
1. Automatickou indexací na základě morfologických pravidelností českého jazyka – typických koncovek (zachycených v souboru dictio.cfg)

Příklad 1: Dle čeho Prospéro identifikoval výraz ***novičok*** jako *entité*, přičemž víme, že dosud výraz ve slovnících uložen takto samostatně nebyl? Ve souboru dictio.cfg je pod *entités* uvedena přípona *-ok*, Prospéro tudíž přiřadil slovo novičok k *entités*:

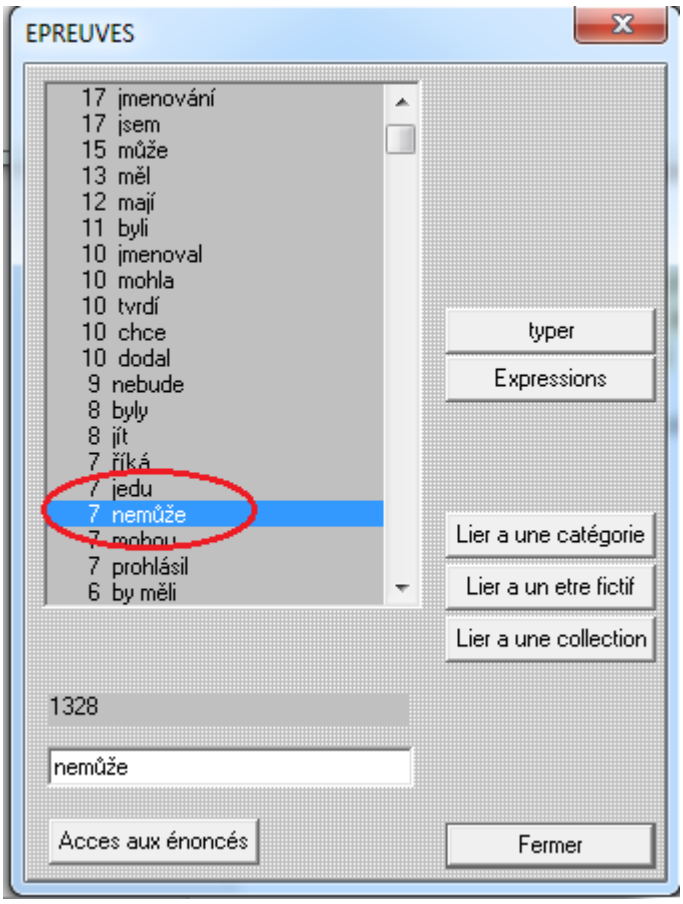


Příklad 2: Prospéro identifikuje výraz ***twitterovém*** jako adjektivum, tentokrát na základě koncovky *ém*, která je uložena ve slovníku dictio.cfg pod *qualités*:

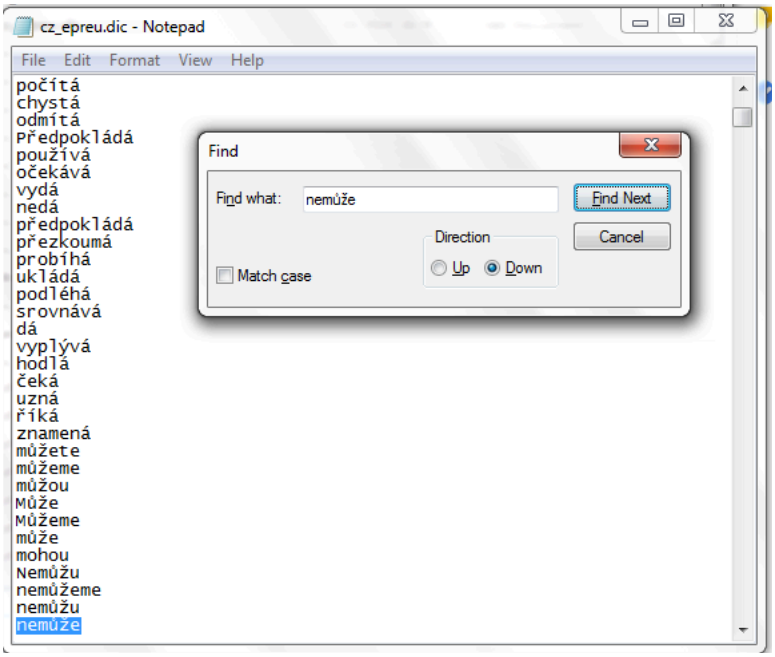


2. Automatickou indexací na základě seznamu již kategorizovaných slov (zachycených v souborech etre.dic, quali.dic, epreu.dic, marqu.dic, nomb.dic, outi.dic, viz výše)

Slovo ***nemůže*** není vyhledáno přes koncovku, ale je ve slovníku zadáno ručně: Prospéro se „podíval“ do slovníků, a protože výraz již byl v příslušném slovníku uložen, klasifikuje ho tak i v našem projektu.



Ukázka struktury slovníku cz_epreu.dic uloženého při instalaci:

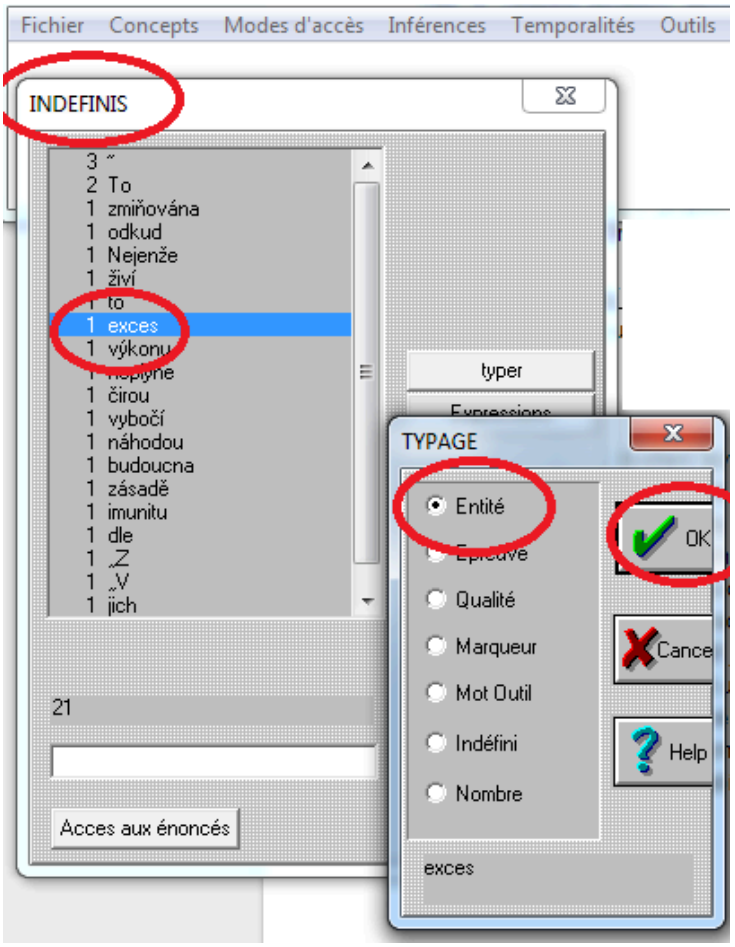


U těchto dvou příkladů *(**novičok, nemůže**)* fakticky není pro naši práci podstatné, dle čeho je Prospéro řadí mezi *entités,* resp. *épreuves*, jde ovšem o to, abychom rozuměli, jak a proč je to které slovo klasifikováno, příp. jak změnit jeho klasifikaci, jak ukládat do slovníků, jak dle potřeby měnit ve slovnících systém přípon a koncovek (neboť není

dokonalý) atd. Pro úpravu, změnu, přesun výrazu do jiné kategorie pak slouží následující bod.

3. Manuální indexací dosud neznámých slov anebo manuální úpravou kategorie slov: funkce *Typer*

Příklad 1: V menu *Fichier – Modes d'accès – Indéfinis* (tedy v seznamu dosud nezařazených slov) vidíme, že slovo ***exces*** není ještě klasifikováno jako *entité*:



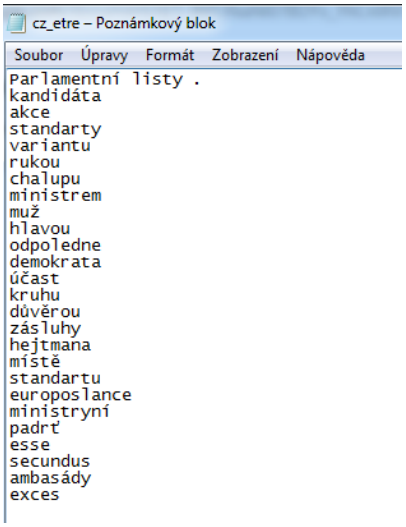
Zařazení slov do *indéfinis* neznamená, že s nimi Prospéro nepracuje. Z hlediska analytických možností popsaných níže (např. formulí) je sice výhodnější zpřesnit typ slov, který chceme hledat, ale můžeme také vždy vytvořit dotaz, který počítá s jakýmkoli typem slov, *indéfinis* nevyjímaje.
Prospéro také mnohdy nepozná záporné podoby sloves, a řadí je tak mezi *indéfinis*; v našem případě např. sloveso ***neplyne***.
Všimněme si rovněž v seznamu *indéfinis* spojení ***„Z*** a ***„V***. Pocházejí z našeho neupraveného textu a ukazují, jak Prospéro zachází s českými uvozovkami. Chybně je považuje za písmena – za počáteční písmena neznámých slov. Jakmile vyměníme české uvozovky za anglické, najdeme správně ***Z*** a ***V*** v seznamu *Mots outils*.

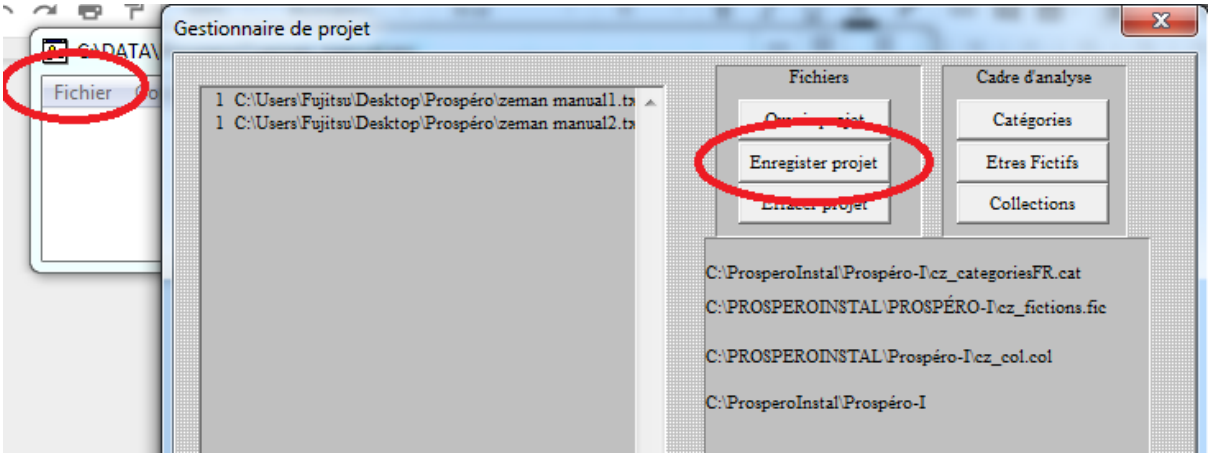
První způsob pro uložení dosud neznámého slova je následující: najedeme na slovo kurzorem, klikneme na tlačítko *Typer*, zvolíme příslušnou kategorii, zde *entité*, a potvrdíme OK.

Správné uložení slova ověříme otevřením souboru cz_etre.dic: nově klasifikované slovo je vždy ihned přidáno na konec seznamu.

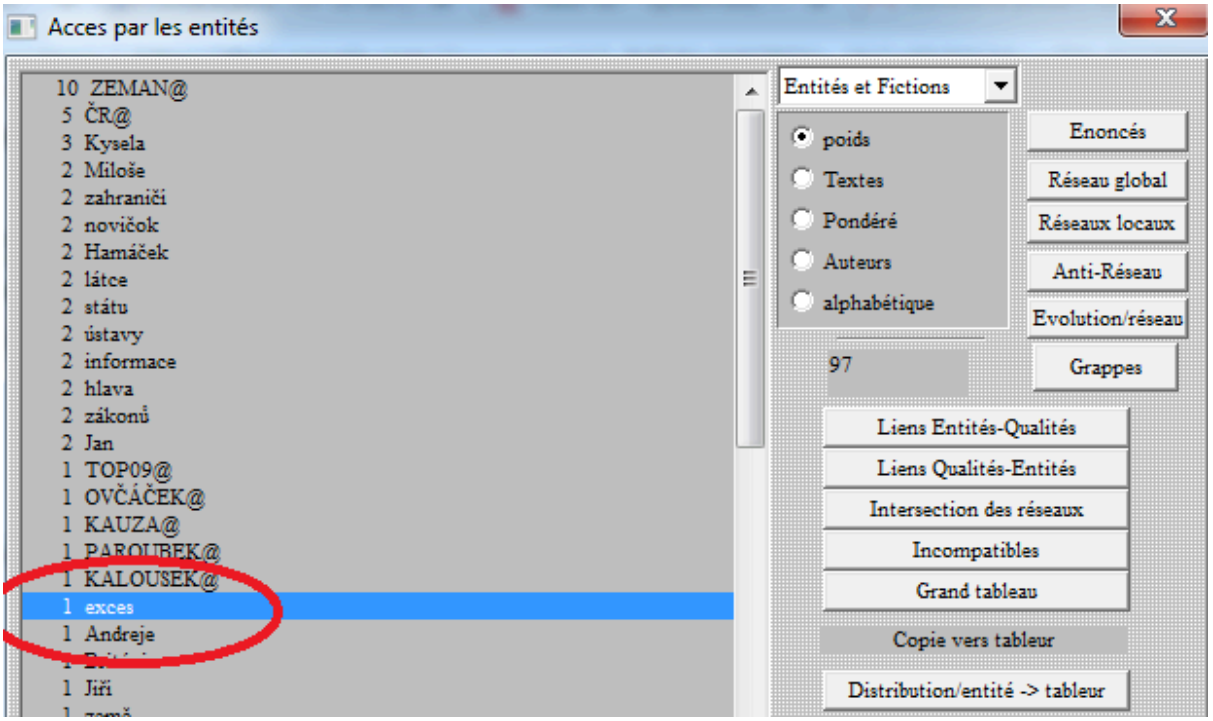
Kromě ukládání v rozhraní Prospéra lze přidávat slova i druhým způsobem, přímo do instalovaných souborů. Chceme-li tedy uložit např. *entité* ***exces*** jiným způsobem, otevřeme soubor cz_etre.dic a připíšeme slovo na konec:

Vždy je dobré volit pouze jeden z těchto způsobů!





Příště už slovo ***exces*** bude vždy klasifikováno jako *entité*. Kontrola po znovunačtení projektu:



Co se děje v Prospéru, když manuálně (pře)klasifikujeme slovo?
Prospéro slovo přiřadí k jednomu z jazykových slovníků, což se děje ihned po kliknutí na tlačítko *Typer*. Pro aktivizaci změny je však třeba projekt znovu načíst – kliknout na tlačítko *Corpus entier* (pod nápisem *Retraitement automatique*) a pak na *Lecture des structures de corpus*. Teď už chápeme, proč je důležité tento postup (viz oddíl 1.5) pravidelně opakovat: jen tak Prospéro vytvoří nový BDT soubor pro každý text v daném projektu, jinak by pracoval s těmi, které jsme vytvořili dosud.
BDT soubor je výsledkem čtení textů na základě jazykových slovníků. Obsahuje seznam slov podle typu a umístění v textu. Otevírá se v Poznámkovém bloku.
Pokud jsme tedy od posledního načtení např. žádné slovo nepřeklasifikovali, stačí projekt otevřít a kliknout jen na *Lire des structures de corpus.* Pro jistotu je však vždy doporučeno kliknout na obě tlačítka.

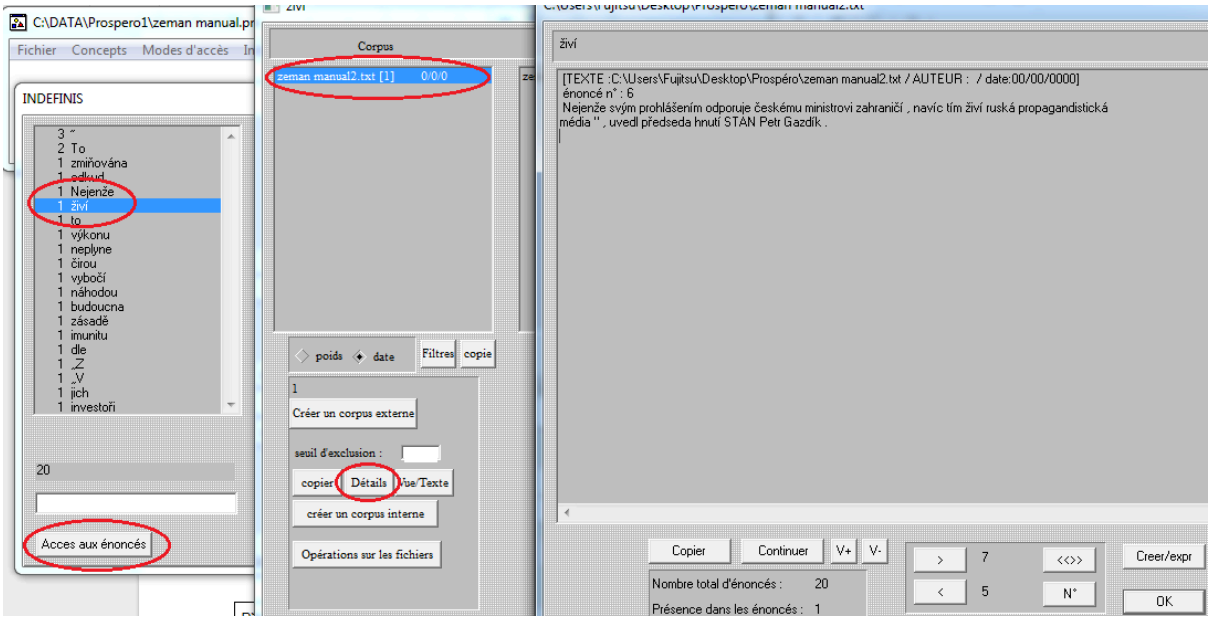
Příklad 2: Potřebujeme změnit klasifikaci slova ***Rusové***, neboť to Prospéro dosud řadil dle homonymní koncovky *-ové* ke *Qualités* (plyne z poměrně časté shody koncovek v češtině):

Prospéro klasifikuje slova začínající velkým písmenem jako *entités*, koncovka má ale vždy před velkým písmenem přednost, proto je tedy slovo ***Rusové*** v adjektivech. Kvůli velkému počátečnímu se zase výraz ***Tudíž*** vyskytuje v *entités*. Buď jej manuálně překlasifikujeme na *marqueur*, nebo změníme manuálně na podobu ***tudíž*** a Prospéro jej pak zařadí správně mezi *marqueurs* sám.

Více k práci s velkými písmeny u vlastních jmen viz oddíl iii. Jména a příjmení.

Příklad 3: Slovo ***říci*** je klasifikováno jako *entité* kvůli koncovce *-ci*, uvedené ve slovníku Dictio pro *entités*. Překlasifikujeme je tedy ručně na sloveso. Postupujeme stejně jako v předchozím případě, tedy pomocí tlačítka *Typer* uložíme do příslušného slovníku.

Příklad 4: U českých textů se poměrně často setkáváme s lexikální polysémií či homonymií. V seznamu *indéfinis* tak kupř. vidíme slovo ***živí.*** Můžeme je uložit buď jako sloveso, nebo jako přídavné jméno. Pro usnadnění rozhodnutí je užitečné podívat se na kontext, v němž se slovo v původním textu vyskytuje. Můžeme postupovat takto: označíme slovo ***živí***, klikneme na *Acces aux énoncés*, označíme daný text a přes *Détails* pak uvidíme konkrétní větu.



Podobně je tomu např. s výrazem ***jedu***, v našem kontextu se může jednat o substantivum ***jed***, či sloveso ***jet*** a je třeba zvážit, v jakém kontextu je pro nás výhodnější mít slovo uložené. Takto lze postupovat pouze jednotlivě, tzn. každé slovo je třeba ukládat zvlášť.

Připomeňme, že zatímco v prvních dvou případech Prospéro sice slova přečte, ale do slovníků neuloží (u prvního slova je to zbytečné díky pravidelnosti koncovky, druhé už tam figuruje), ve třetím případě je do slovníků naopak sami aktivně ukládáme.

Další způsoby klasifikace a překlasifikace slov budou popsány níže v kapitolách Kategorie, Vyhledávání a Formule**.**

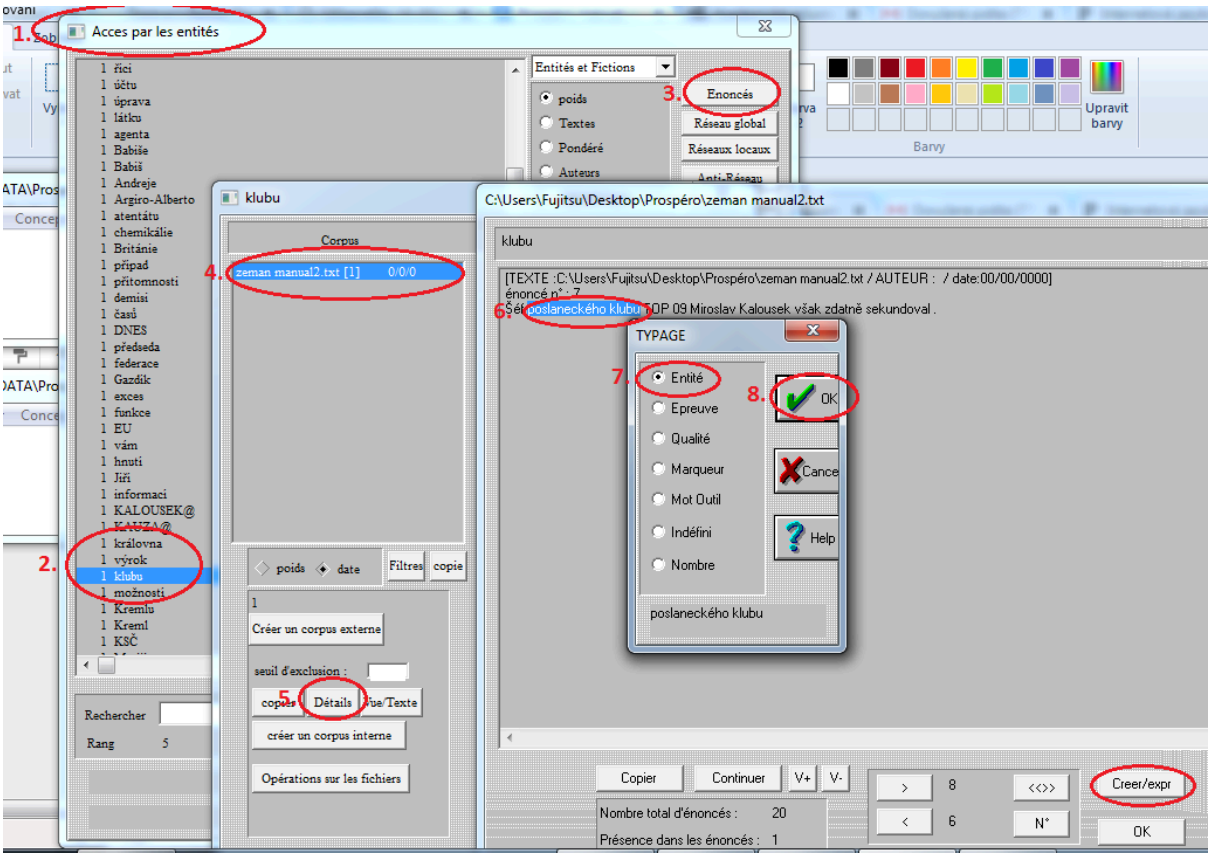
ii. Vytváření frazémů

Tematicky patří pod bod č. 3 i **vytváření frazémů** (ustálená spojení minimálně dvou slovních tvarů, *expressions*). Uživatel může doplňovat slovníky víceslovnými výrazy, pokud chce, aby je Prospéro četl jako jeden výraz dohromady (jako jedno slovo). Některé frazémy jsou ustálené, jednoznačné a přenosné mezi korpusy (***kout pikle, chytat lelky, mít fortel***), některé jsou ovšem specifické vzhledem k odborným tématům. Např. slovo ***hromada*** (nevyskytuje se v našich textech) se tak může vyskytovat v primárním významu (množství něčeho na sobě nakupeného), v případě tematicky odpovídajícího korpusu pak ale spíš v dalším významu schůze/shromáždění, tedy v podobě ***valná hromada***. Pokud se tento význam v korpusu opakuje a je pro nás výhodné, aby Prospéro slova ***valná*** a ***hromada*** klasifikoval dohromady jako jednu *entité*, vytvoříme tzv. *expression*. Nevyloučíme tím ovšem možnost, aby Prospéro přečetl samostatně slova ***hromada*** i ***valná*** pokud se vyskytují nezávisle od sebe (např. ve spojení ***valná část, hromada slov***). Frazémy jsou uloženy zvlášť pro každý zadaný tvar (***tajné služby, tajných služeb*** atd.). Výhodné je to např. pro sousloví ***MF Dnes/Mladá fronta Dnes, Ruská federace*** aj.

Při vytváření frazémů je důležité zvážit, zda jimi doplnit základní slovníky, nebo vytvořit kopii slovníků za účelem zkoumání pouze daného tématu/korpusu, pokud má velmi zvláštní terminologii. Vytvořením frazémů se totiž změní výpočty a inference, které Prospéro provádí.

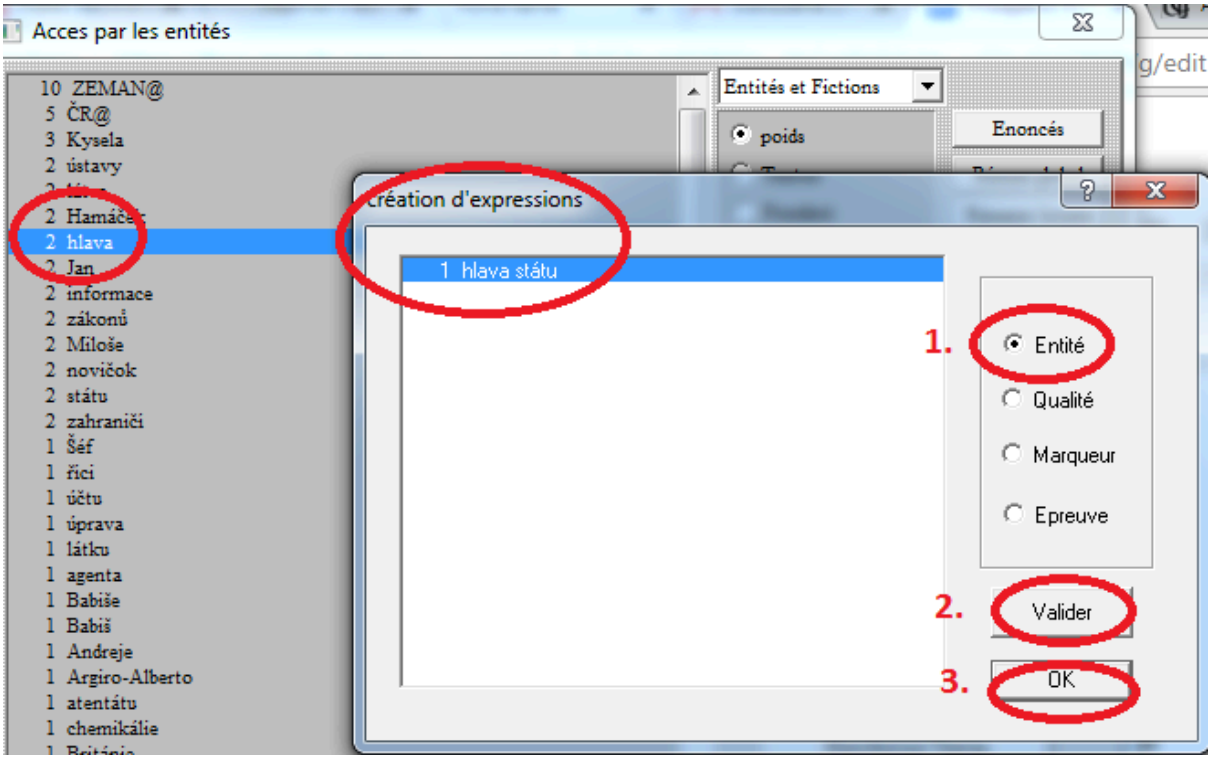
Frazémy lze vytvářet několika způsoby:

1. z libovolné věty, kterou zobrazíme v okně pro čtení textů (po kliknutí na *Détails* anebo *Accès aux textes*), dle návodu; zde např. sousloví ***poslaneckého klubu***:

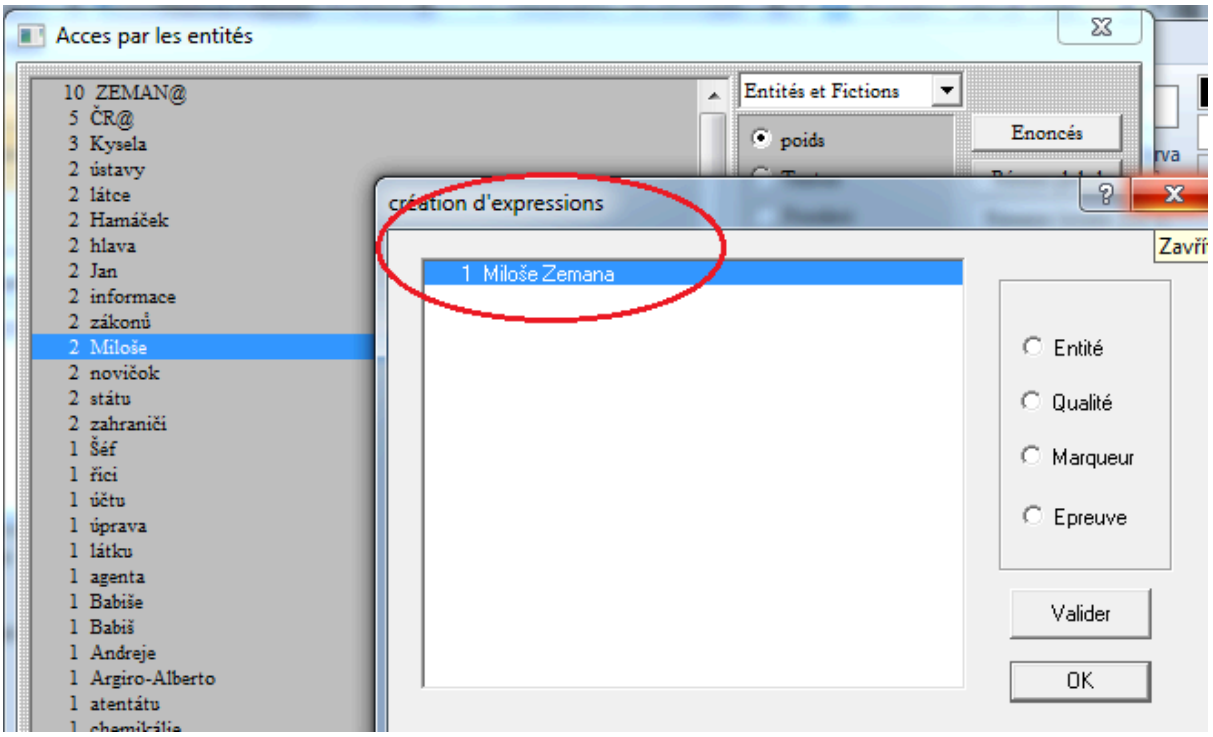


2. v seznamu *entités* existuje možnost vidět některá slova v kontextu jejich nejčastějších kotextů (frekventovaných spojení, např. pravidelně se opakující přívlastek s podstatným jménem). Kliknutím na *expressions* lze tato spojení vybrat a „typovat“, tj. uložit jako *expression* (tato možnost ovšem nefunguje pro všechna slova).

Příklad: Vidíme v seznamu *entités* slovo hlava a tušíme, že může být součástí frazému. Když kurzorem označíme slovo ***hlava*** a vpravo v menu klikneme na tlačítko *expressions*, v nově nabídnutém okně pak kurzorem zjistíme, že se slovo vyskytlo (jednou) ve spojení ***hlava státu***. Abychom uložili tento výraz jako frazém, označíme ho a zvolíme v tomto případě *entité*, pak klikneme na *Valider* a OK. Odteď bude spojení ***hlava státu*** Prospérem klasifikováno jako jedno slovo náležející mezi *entités*. U ostatních slovních druhů postupujeme analogicky.

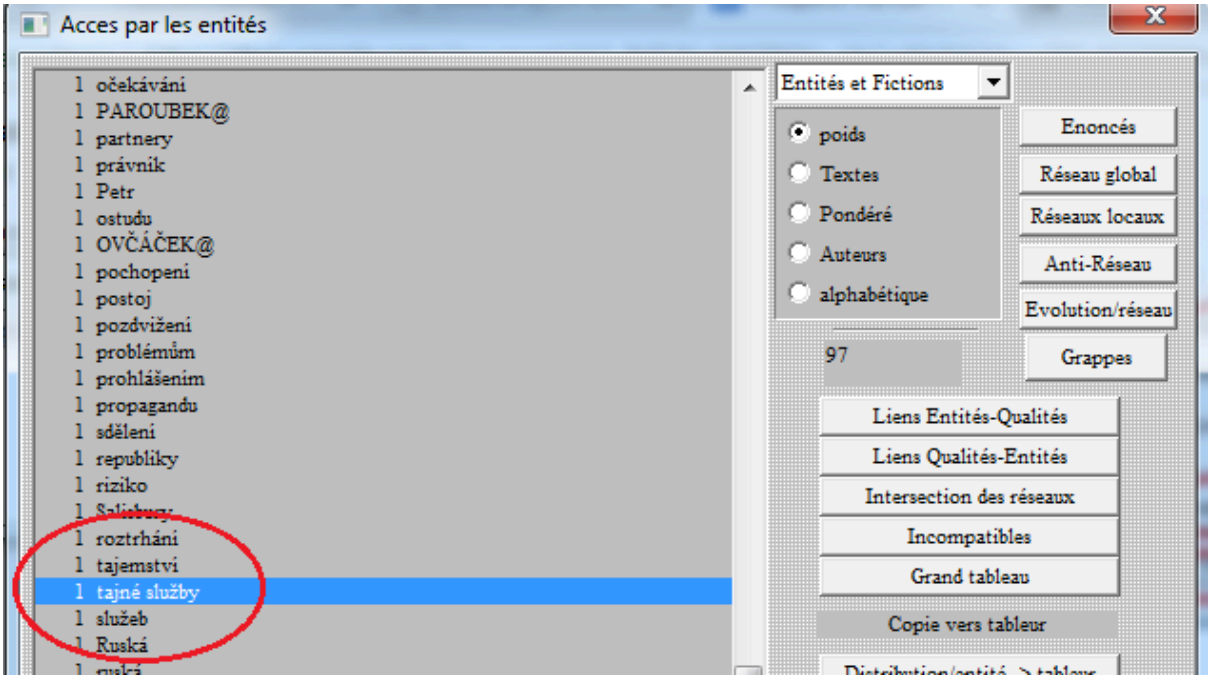


Takto v našem vzorovém korpusu Prospéro vnímá (navrhuje) jako *expression* také např. sousloví ***Miloše Zemana***:



Jména osob ovšem neukládáme jako *expressions*: pro zacházení s nimi má Prospéro zvláštní sadu funkcí (viz následující oddíl Jména a příjmení).

Některé frazémy, např. sousloví ***tajné služby***, jsou již v našem projektu jako *expressions* uloženy:



Další způsoby vytváření a využívání frazémů budou popsány v příslušných kapitolách.

iii. Jména a příjmení

Rozpoznání jmen a problém jejich skloňování

Prospéro identifikuje vlastní jména na základě seznamu křestních jmen (soubor prenoms.dic). Většina běžných českých jmen už v něm figuruje. Potom umí spočítat postavy korpusů podle vzoru: křestní jméno + slovo začínající velkým písmenem. Pokud se v korpusu minimálně jednou vyskytlo příjmení po jménu uvedeném v prenoms.dic, spojuje i izolované výskyty toho příjmení s danou osobou. Takže pokud Prospéro už zná spojení ***Miloš Zeman***, ví taky, že slovo ***Zeman*** odkazuje na stejnou osobu. Určitou komplikaci představuje skloňování, a to ze dvou důvodů. Za prvé, aby Prospéro správně identifikoval jména našeho korpusu, musel by soubor prenoms.dic obsahovat všechny pády českých jmen, a v tomto stádiu tomu tak zdaleka není. Za druhé, Prospéro bere každý pád jako jinou osobu. Můžeme tedy přidat do souboru prenoms.dic tvar ***Miloše*** a Prospéro pak rozpozná spojení ***Miloše Zemana*** jako jméno osoby, ale nemá propojení tohoto tvaru se jménem ***Miloš Zeman***, a bude tedy vést samostatné seznamy obou „osob“ (a další pro pádové tvary ***Miloši Zemanovi***, ***Milošem Zemanem*** atp.). Tím pádem by bylo užitečné doplnění souboru prenoms.dic o všechny pády českých jmen.

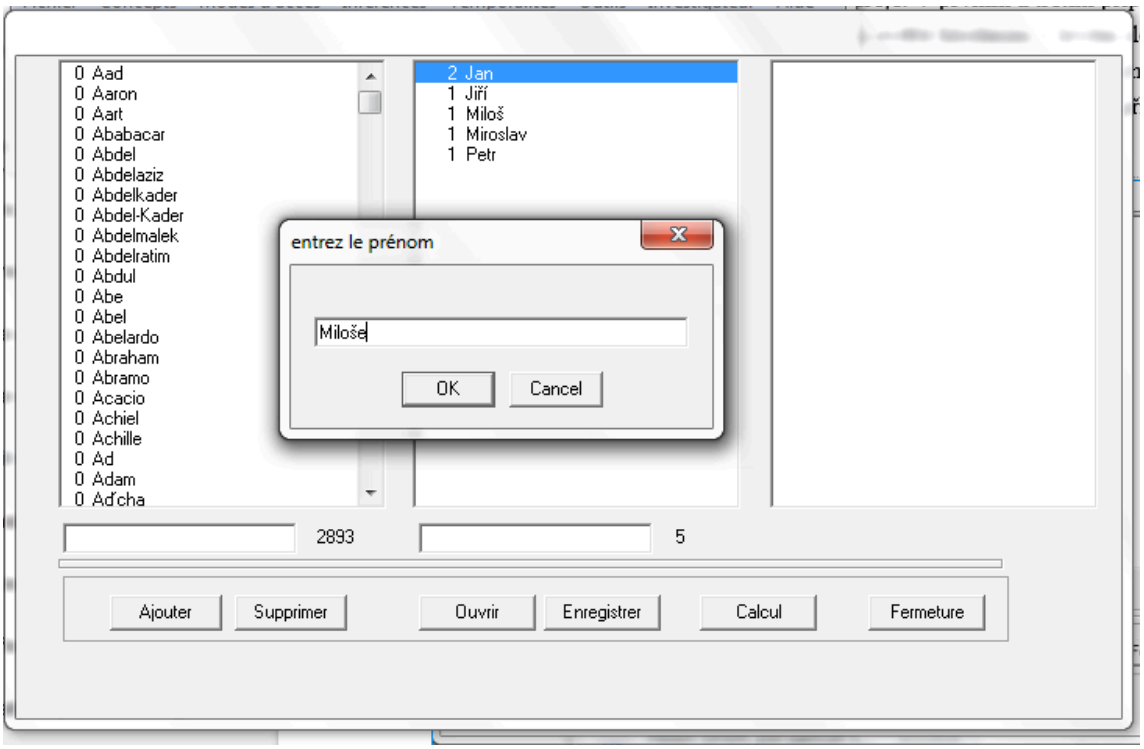
Soubor dictio.cfg zase obsahuje běžná česká oslovení p./pan/paní (ve všech pádech) pod heslem [INTRO_PERSONNE], aby Prospéro věděl, že pokud následující slovo začíná velkým písmenem, označuje osobu, a aby je uměl spojit se stejným příjmením vyskytujícím se samostatně anebo po křestním jménu.

Prospéro bohužel neidentifikuje české znaky s diakritikou psané velkými písmeny. Pro lepší validitu výsledků je proto žádoucí přepsat počáteční velká písmena příjmení, konkrétně Ž, Š, Č, Ř, Ď, Ť, Ň, Á, Ú, Ú, Ó, É, Í bez diakritiky (např. Eva Černá se změní na Eva Cerná).

Identifikace osob v korpusu

Pro použití této funkce je nejprve třeba specifikovat adresu souboru prenoms.dic v záložce *Fichier – Configuration du projet.* Příkazem *Modes d'accès – Listes – Patronymes – Prénoms* potom získáme přehled jmen v našem korpusu. V levém sloupci jsou veškerá jména, která Prospéro zná (pokud se nic neobjeví, je třeba ověřit, zda je cesta k souboru prenoms.dic nastavena správně), ve středním pak všechna jména v korpusu. Pokud zde klikneme na jakýkoli záznam, v pravém sloupci se objeví všechna příjmení spojená s tímto jménem. V tomto okně (tlačítko *Ajouter* a pak *Enregistrer*) anebo přímo do souboru prenoms.dic můžeme přidat jména, která v paměti Prospéra chybějí.

Pokud porovnáme (od oka) seznam jmen v prostředním sloupci s texty našeho cvičného korpusu, je vidět, že tři jména zmíněná v druhém textu zde chybějí: ***Miloše***, ***Mariji*** a ***Andreje***. V prvním a třetím případě jde o známý problém s českým skloňováním, protože si můžeme snadno ověřit hledáním v levém sloupci, že zatímco ***Miloš*** a ***Andrej*** se nacházejí v seznamu jmen, která Prospéro zná, jejich nenominativní pády v paměti nejsou. ***Marija (Mariji)*** tam ale chybí zcela. Chceme-li je přidat, klikneme na *Ajouter* a vyplníme nabízené pole (*entrez le prénom*).



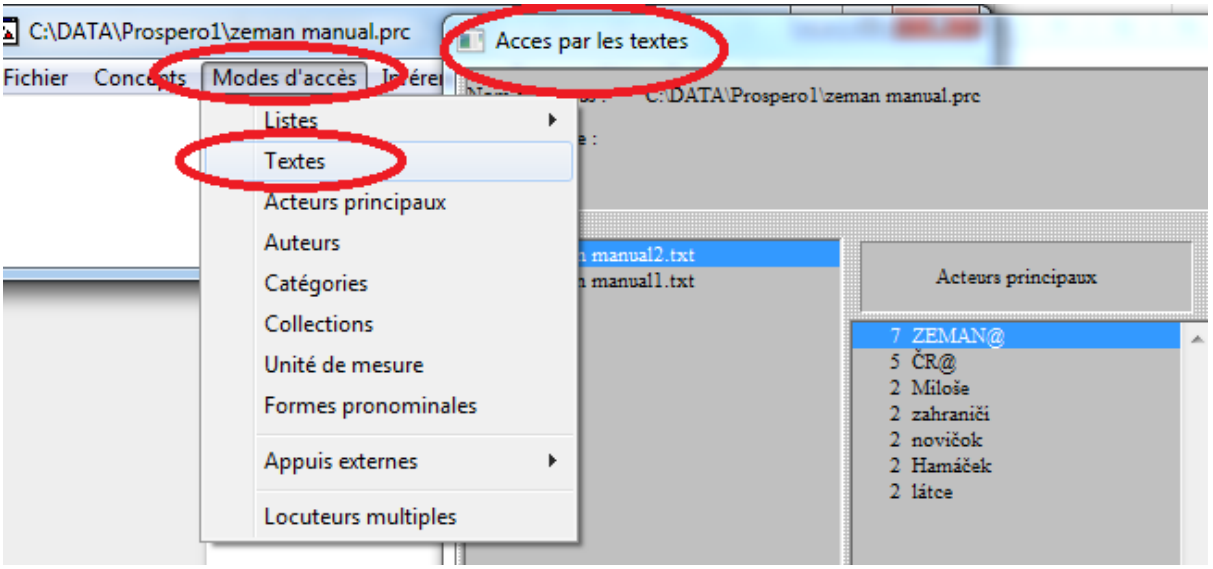
Spuštěním příkazu *Modes d'accès – Listes – Patronymes – Noms* můžeme získat nejen přehled příjmení v korpusu, ale i seznam vět, ve kterých je daná osoba zmíněna. Klikneme-li na příjmení, zjistíme, jak je tato osoba „představena“ (například křestním jménem, oslovením pan/paní/p., případně akademickým titulem) a kolikrát se v korpusu vyskytuje, a chceme-li vidět dané kombinace jmen/oslovení + příjmení v kontextu vět, klikneme na *Enoncés (nom+prénom)*. Podobně tlačítko *Enoncés (nom)* poskytuje přístup ke všem textům, kde je daná osoba zmíněna bez ohledu na způsob označení. Dva sloupce ve spodní části okna umožňují začít zkoumat vztahy osob v textech se vyskytujících jednak s dalšími osobami (*Personnes physiques du réseau*), jednak s dalšími kapitalizovanými podstatnými jmény (čili zejména s institucemi, organizacemi) (*Autres noms propres du réseau*).

V jakémkoli sloupci tohoto okna – a to platí ve většině oken Prospéra – můžeme kopírovat celý seznam nebo jeho horní část stisknutím pravého tlačítka myši a volbou *copier liste*. V dialogovém okně specifikujeme počet záznamů (odshora), které chceme kopírovat. Pokud máme zájem o celý seznam, stačí zadat jakýkoliv počet převyšující odhadovaný počet položek v našem seznamu (máme-li tedy např. 50 položek, zadáme klidně 100, abychom s jistotou obsáhli všechny).

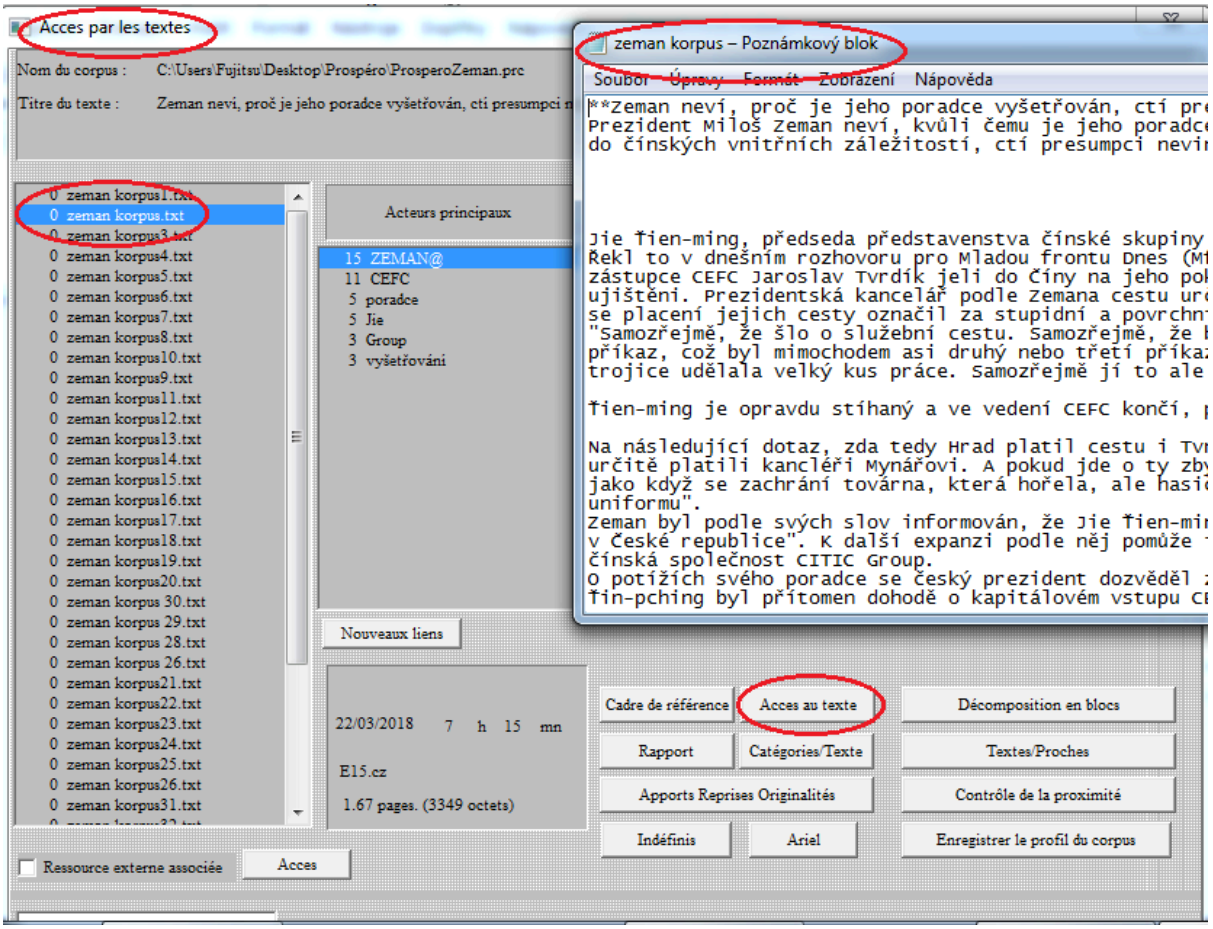
b. Základní orientace v textech

(Menu *mode d'accès – textes*)

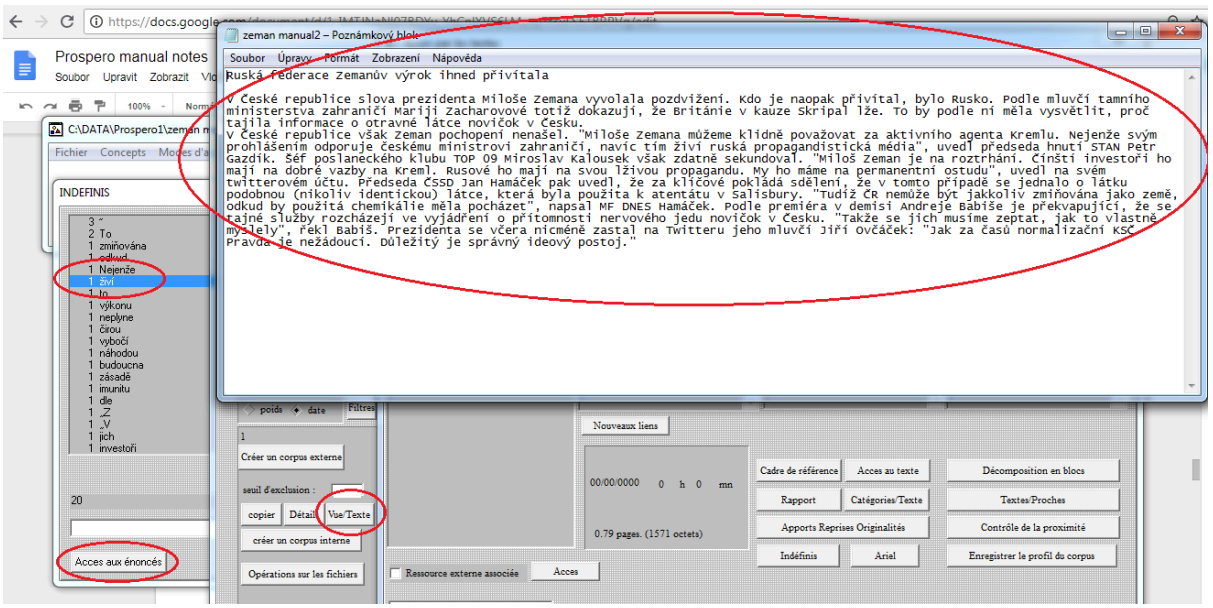
Do rozhraní pro texty, s nímž budeme pracovat v této kapitole, přistupujeme z hlavního menu Prospéra takto:



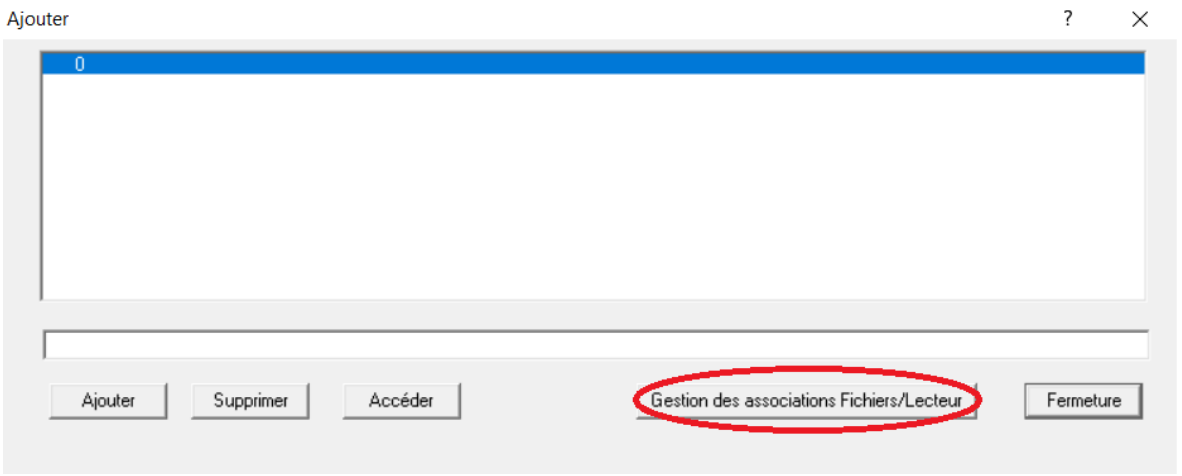
Postup zobrazení konkrétního textu ve formátu NotePadu je pak přes Menu Modes d'acces – Textes (označíme myší text) – Acces aux textes



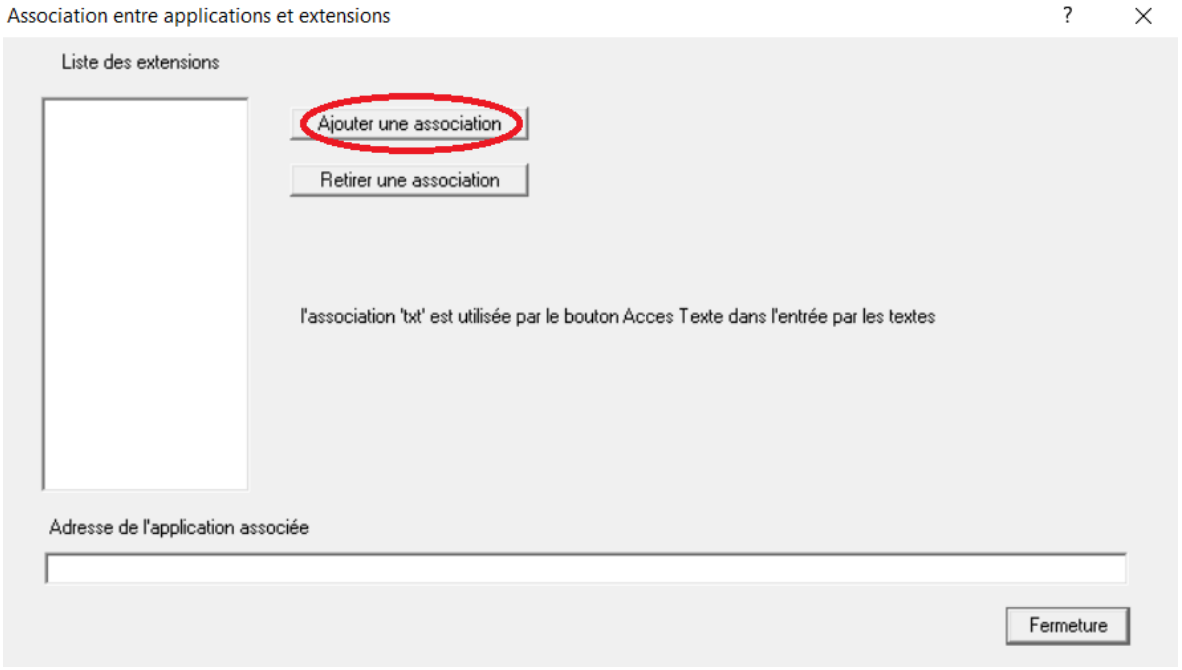
Máme také možnost otevřít text přes určitou vyhledanou jazykovou jednotku. Jako příklad použijeme opět slovo ***živí***, označíme je, klikneme na *Acces aux énoncés*, označíme daný text a přes *Vue/Texte* si opět zobrazíme text v Poznámkovém bloku:



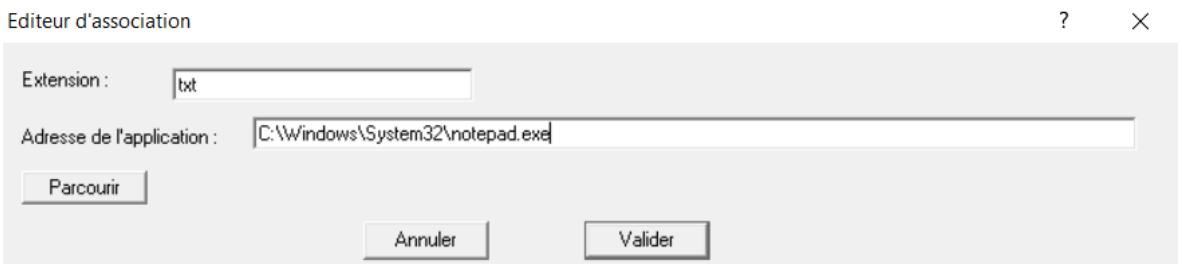
Pokud se Poznámkový blok neotvírá, je třeba postupovat následovně: *Mode d'accès* > *Textes* > *Access* > *Gestion des Associations... Ajouter une Association* '*txt*'. Zde zvolíme formát .txt a na řádek *Adresse de l'application* doplníme adresu spouštěcího souboru Notepad.exe – C:\Windows\System32\notepad.exe



V následujícím okně klikneme na tlačítko *Gestion des associations Fichiers/Lecteur*, které nás zavede do hlavního okna správy asociací.

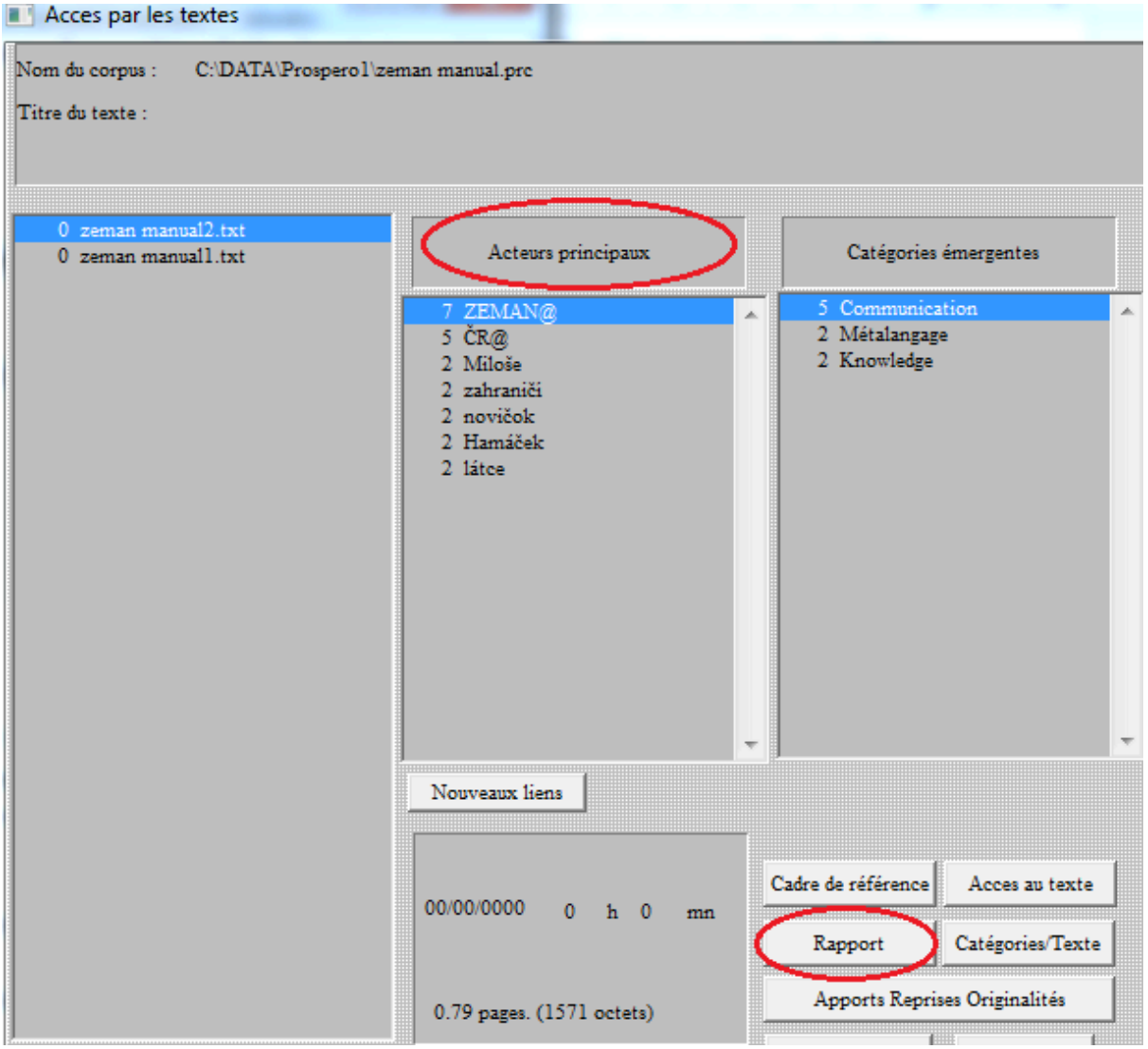


Nejprve je nutno zadat cestu k textovému editoru. Tuto cestu nastavíme tlačítkem *Ajouter une association*. Druhé tlačítko *Retirer une association* (odstranění asociace) využijeme v případě, kdybychom udělali během nastavení chybu.



V následující záložce nejprve vyplníme typ přípony txt na řádku *Extension*.
Na dalším řádku se pak nastavuje konkrétní cesta (adresa spouštěcího souboru) k textovému editoru. Tato cesta se může na konkrétních osobních počítačích a instalacích operačního systému Windows lišit. V našem případě se jedná o C:\Windows\System32\notepad.exe.
Nejprve doporučujeme ověřit, že se jedná o skutečné umístění spouštěcího souboru Poznámkového bloku.
Stisknutím tlačítka *Valider* (potvrdit) je cesta nastavena.
Doporučení: U těchto typů nastavení doporučujeme znovu uložit celý projekt (v hlavním menu *Fichier – Gestion du projet – Enregistrer projet*). Bez uložení projektu by bylo třeba cestu nastavovat při každém „načtení“ projektu.

V nabídnutém rozhraní vidíme dále např. seznam tzv. *Acteurs principaux* či funkci *Rapport*:



Acteurs principaux jsou hlavní postavy textu, tj. nejčastěji zastoupené entity vyskytující se v textu; Prospéro je vyhodnocuje sám na základě četnosti užití vzhledem k rozsahu textů a dalších kritérií, tato není možné nijak ovlivnit manuálně.

Užitečný příkaz *rapport* umožňuje zjistit procento neznámých výrazů (*indéfinis*) v našem korpusu. V případě našeho cvičného projektu se jedná o velmi nízké procento, při momentálním stavu slovníků je běžné identifikovat po prvním načtení zhruba od 5 do 15 % neznámých výrazů; hodnotu je třeba dále snižovat klasifikováním neznámých výrazů do příslušných skupin, viz výše. V případě, že by číslo bylo nápadně vyšší, je třeba zkontrolovat cesty ke slovníkům a nastavený jazyk projektu, neboť by se mohlo jednat o instalační chybu. *Rapport* ovšem poskytuje více než jen ukazatel poměru neznámých výrazů: poskytuje stručné shrnutí nejčastěji se vyskytujících pojmů v textu (viz následující oddíl) v porovnání se zbytkem korpusu a na základě hustoty zastoupení konceptů vyzdvihuje několik klíčových vět. Je generován pouze ve francouzštině.

Toto rozhraní nabízí mnohé další funkce, které probereme v příslušných částech výkladu níže.

3. Socio-diskurzivní analýza

Druhá úroveň klasifikace nebo indexace, kterou Prospéro umožňuje, je třídění obsahu korpusu do/dle tzv. **kategorií**, **seznamů/kolekcí** a **velkých postav korpusu.** To je zároveň naší vstupní bránou k interpretační a analytické práci.

Poznámka: Aby bylo možné lépe ukázat pokročilejší funkce Prospéra, a tím se zároveň dostat k hlubšímu pohledu na zkoumané jevy, počínaje následující částí budou již další příklady popsány na rozsáhlejším vzorku textů. Nepůjde zde už tolik o znázornění jednotlivostí jako spíše o práci s vyhledáváním ve větším objemu dat. Všechny texty a jejich metadata (soubory) k vytvoření tohoto cvičného korpusu by měly být dostupné na vašem disku s instalací Prospéra.

a. Pojmy-koncepty a pojmové slovníky: kategorie, kolekce, velké postavy

i. Kategorie

Co jsou kategorie v Prospéru?

Práce s kategoriemi představuje základní klasifikační operaci sloužící klasickému sociologickému kódování, jak je známe z obsahové nebo diskurzivní analýzy. Na rozdíl od kódů jsou však prospérovské kategorie více nezávislé vzhledem k textům, korpusům a jednotlivým případovým studiím: jsou do značné míry mezi korpusy přenosné a univerzálněji použitelné. Není tedy třeba budovat zcela novou sadu kategorií pro každý nový korpus: začneme s aktuální, „standardní“ sadou a postupně, iterativně ji upravíme v průběhu interpretace (měli bychom samozřejmě zachovat někde na bezpečném místě počítače kopii „standardních“ pojmových slovníků a uložit ty, které jsme upravili pro daný korpus, pod novým názvem).

Důležité je pochopit strukturu a logiku kategorií, neboť představují pojmové universum, které stojí v reprezentativním vztahu s diskurzem korpusu – vztah, který by měl být na počátku i v průběhu práce s korpusem znovu promýšlen a upravován. Kategorie mají každopádně tvořit koherentní seskupení pojmů související s vlastnostmi našich korpusů. Stojí například za to pečlivě přečíst text, v němž je určitá kategorie dominantní, za účelem jejího doplnění o případné další, ještě nezachycené prvky téže kategorie. Kategorie totiž často mívají své „prototypické“ texty.

Pozor: je-li zastoupení kategorie v korpusu sice silné, ovšem s jediným slovem v popředí, často to znamená, že v tomto korpusu má kategorie specifičtější nebo úplně jiný význam, než jak byla koncipována (na základě „kódování“ jiných korpusů). V našem korpusu se například kategorie *entités* “*Sanction”* vyskytuje 11krát, ale 6 výskytů se týká slova „ocenění“, v souvislosti s polemikou kolem návrhů na udělení státních vyznamenání. Význam je sice správný, ale mnohem konkrétnější než diskurzivně-narativní význam, který měla kategorie zachytit.

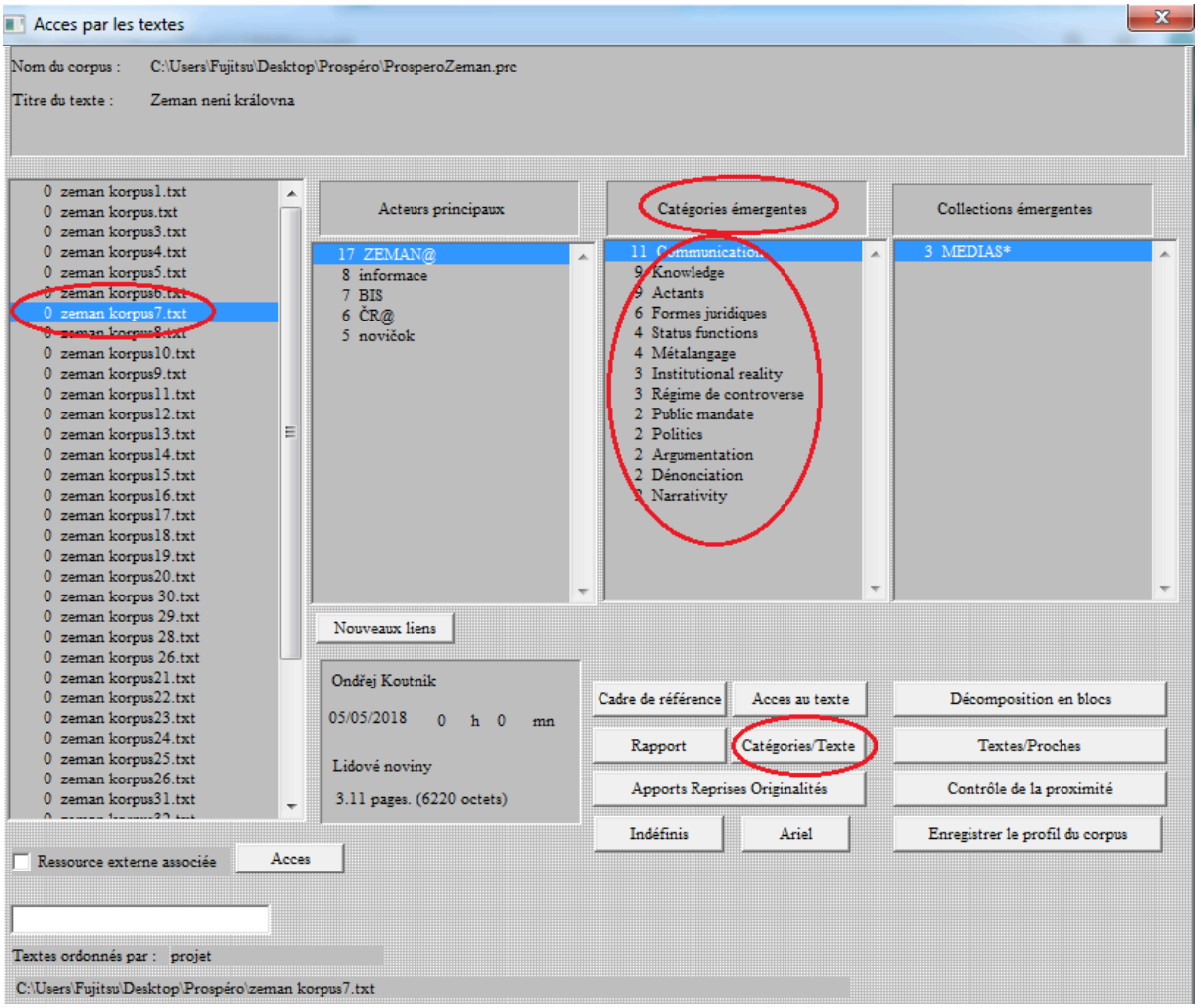
V (mnoha) kategoriích se pohybujeme spíše na úrovni argumentace než na úrovni věcného referování. Například slovo ***přeběhlictví*** je zařazeno do kategorie *Dénonciation* (raději než do kategorie *Politics*), protože bezpochyby odkazuje na sdělení, kterým mluvčí něco nebo někoho odsuzuje.

Práce s kategoriemi

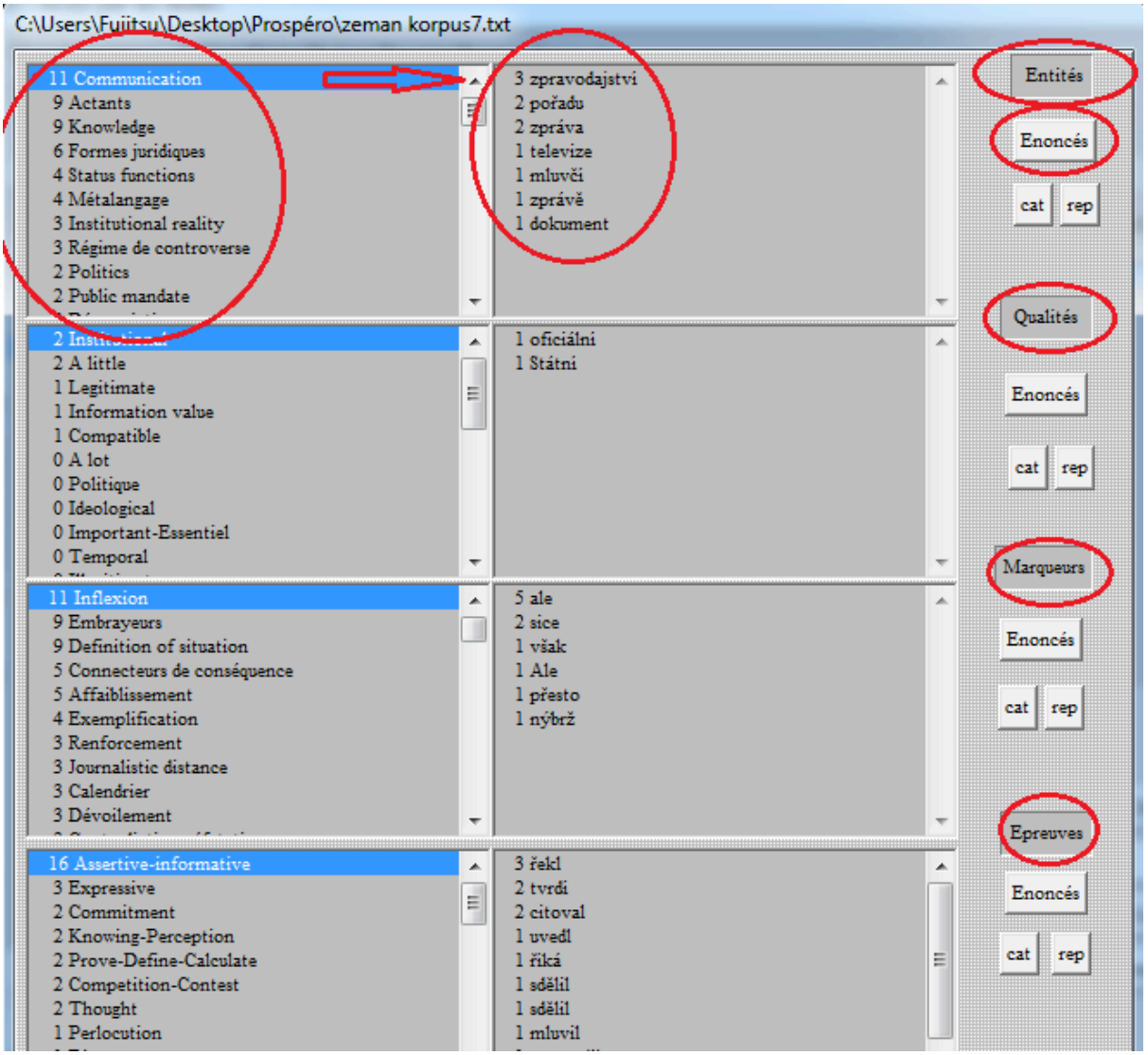
Rozhraní, jejichž prostřednictvím můžeme pracovat s kategoriemi, jsou následující:

A. V rozhraní *Modes d’accès – Textes* vidíme po označení kurzorem v prostředním sloupci *Catégories émergentes* seznam kategorií, jejichž prvky se v daném textu

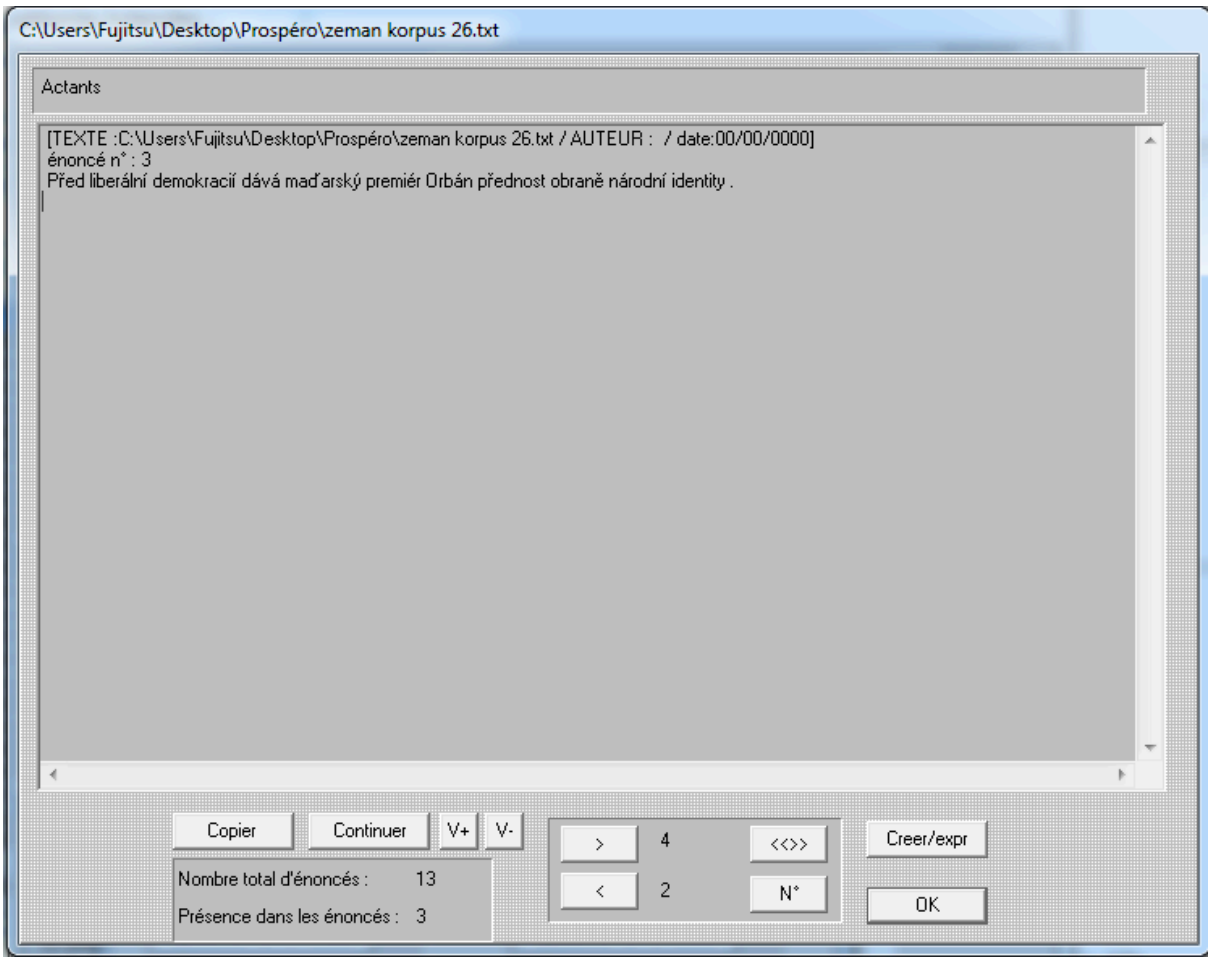
vyskytují častěji:



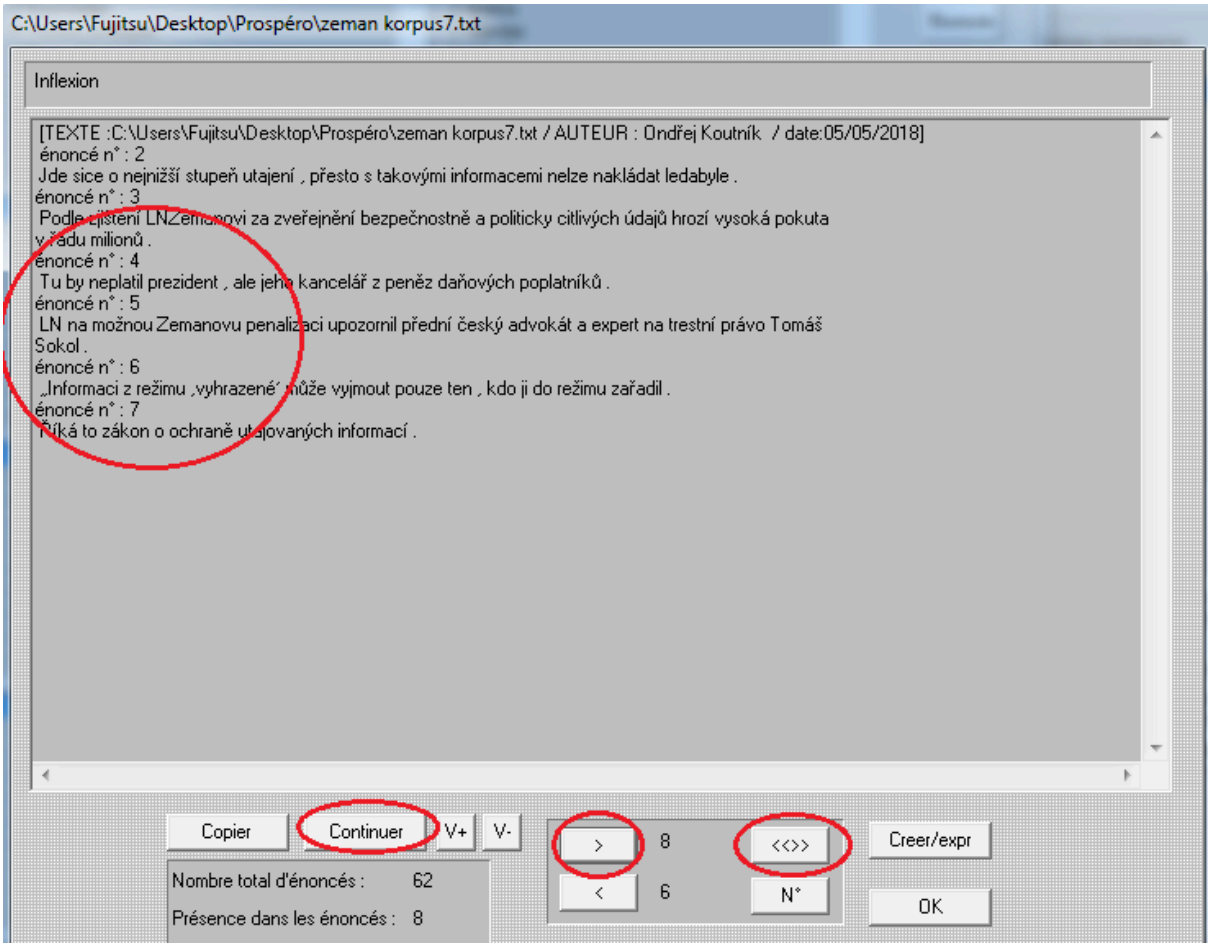
Tlačítkem *Catégories/Texte* pak zobrazíme podrobný rozpis slovních druhů, seřazený dle početního zastoupení:



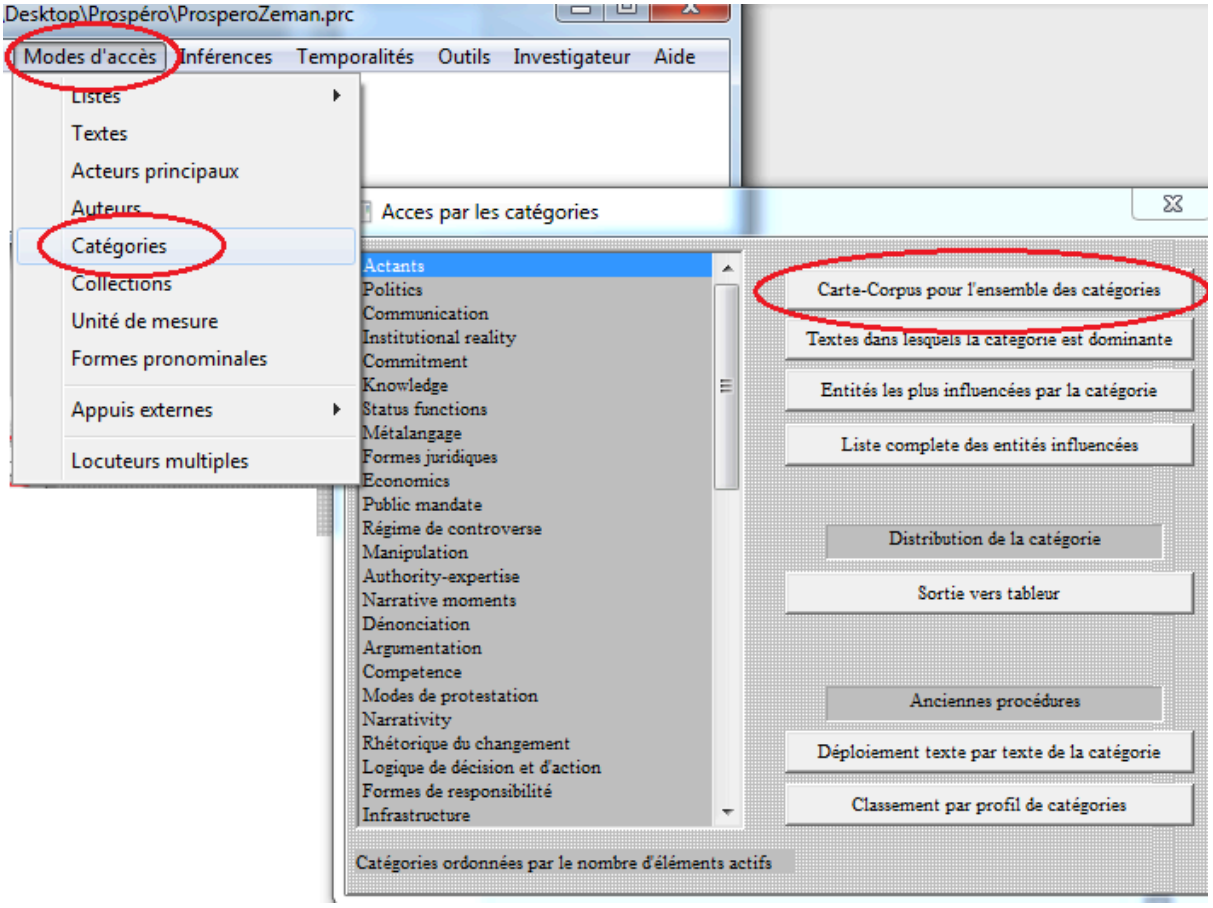
Kliknutím na *Enoncés* se opět dostaneme k celé větě, v níž se daná kategorie či její prvky manifestují:



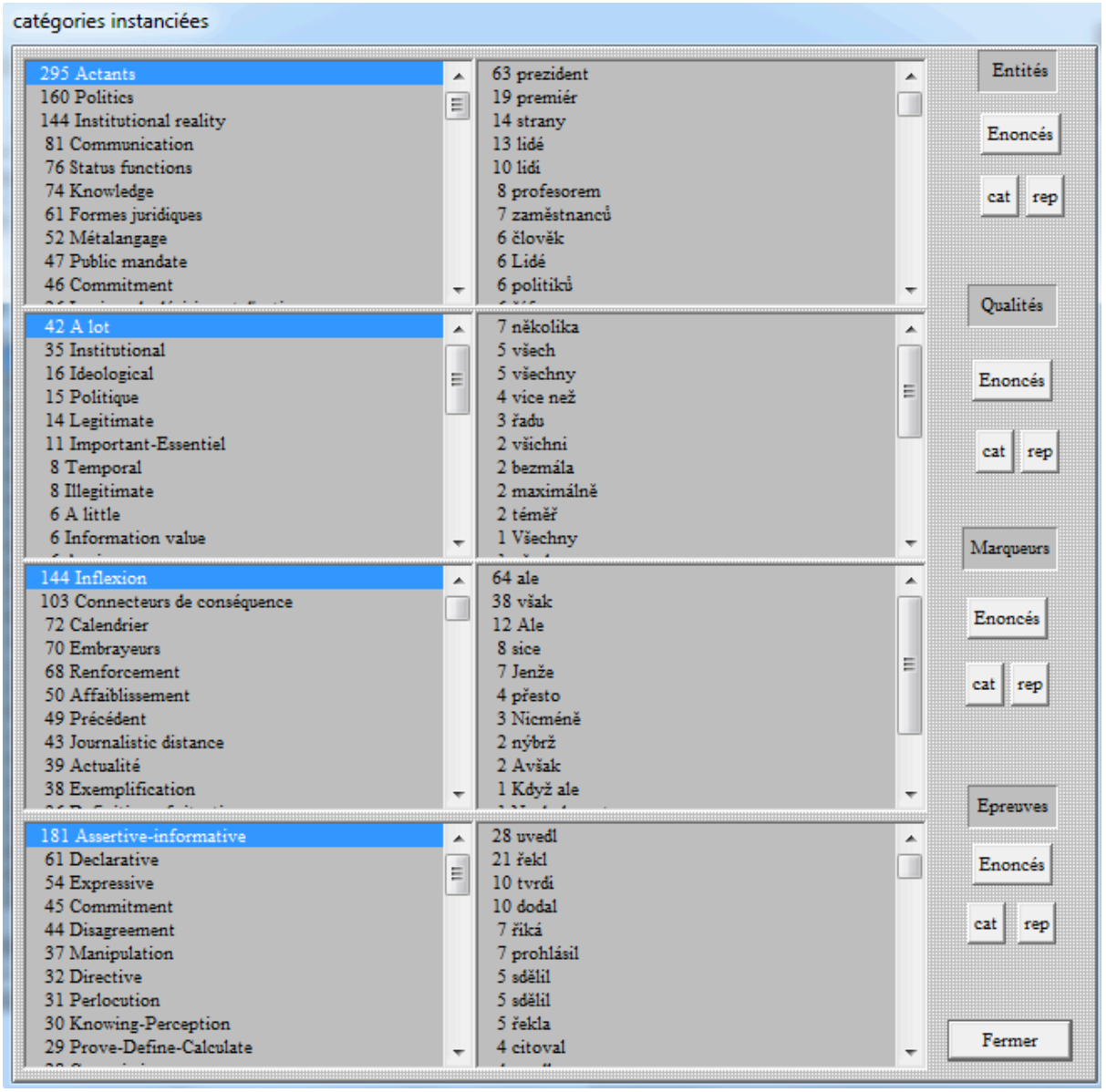
Zde ještě upozorníme na další funkce: kliknutím na šipku doprava zobrazíme následující větu a šipkou doleva předchozí větu (tím nám Prospéro umožňuje číst texty odzadu, což je velmi užitečné pro analýzu argumentace). Po kliknutí na dvojité šipky se objeví výskyt všech vyhledaných zástupců kategorie a na tlačítko *Continuer* další výskyty vybraného slova (jeden po druhém). Tlačítkem *Creer/expr* můžeme vytvořit frazém, viz kapitola výše.



B. Druhou možností je přístup přes *Modes d'accès – Catégories* a tlačítko *Carte-corpus pour l'ensemble des catégories*:

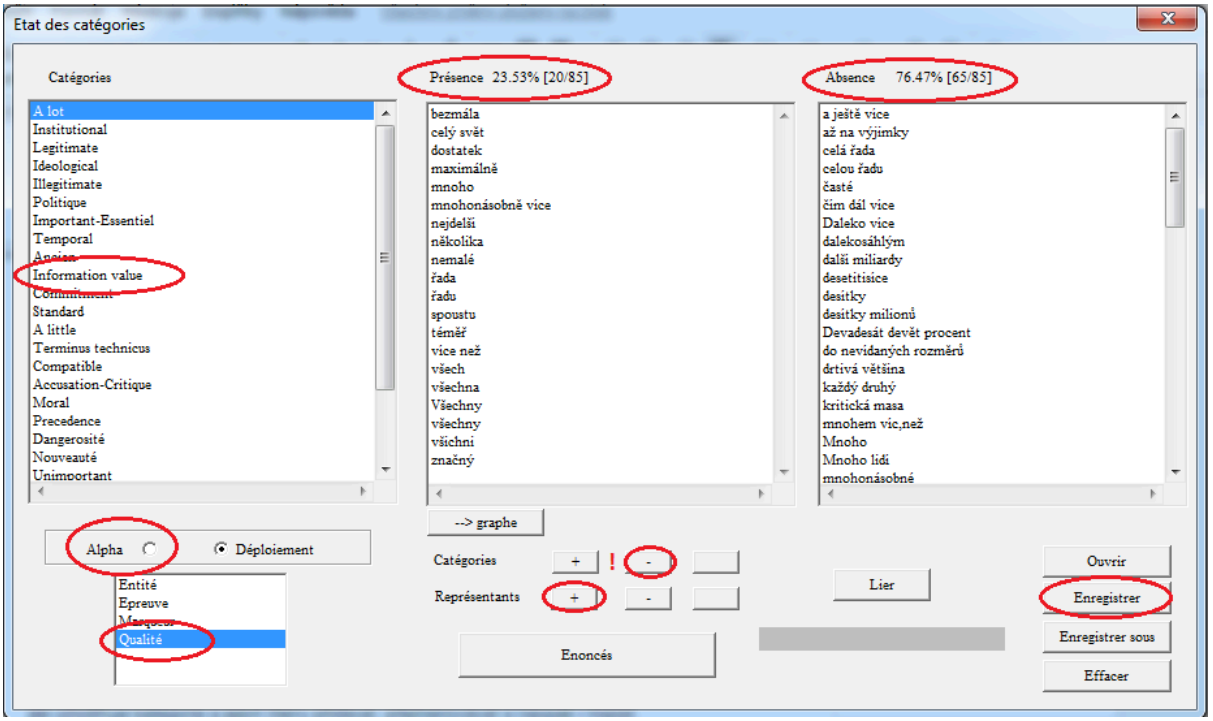


Toto zobrazení má tu výhodu, že umožňuje přehledně zobrazit všechny čtyři skupiny kategorií i s obsaženými výrazy (rozhraní je stejné jako to, které jsme viděli pro jednotlivé texty po kliknutí na tlačítko *Catégories/Texte*):



Pomocí tlačítka *Enoncés* a *Détails* lze opět získat celý kontext zvolené kategorie či slova.

C. Dalším způsobem je přístup přes základní menu: *Concepts – Catégories*:

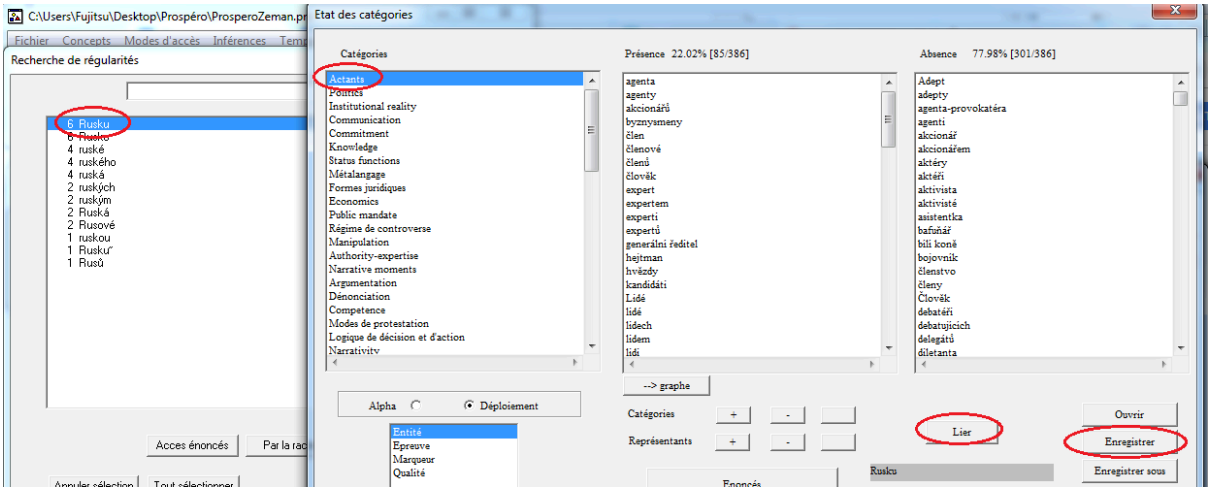


Toto rozhraní poskytuje některé odlišné funkce. Kromě zobrazení seznamu všech kategorií i s jejich obsahem vycházejícím z našeho korpusu ale navíc umožňuje i kategorie a jejich prvky přidávat, přejmenovávat nebo naopak mazat.

Sloupec *Présence* zahrnuje prvky kategorie přítomné v našem korpusu, *Absence* pak ty, které kategorie obsahuje, ale v našem korpusu se dosud nevyskytují. Nové slovo manuálně přidáváme s pomocí tlačítek *Représentants* a *Enregistrer* (uložit). Pozor na kombinaci tlačítka *Catégories* a znaku mínus, jíž bychom odstranili celou kategorii (po uložení). Funkce *Alpha* umožňuje řadit kategorie abecedně.

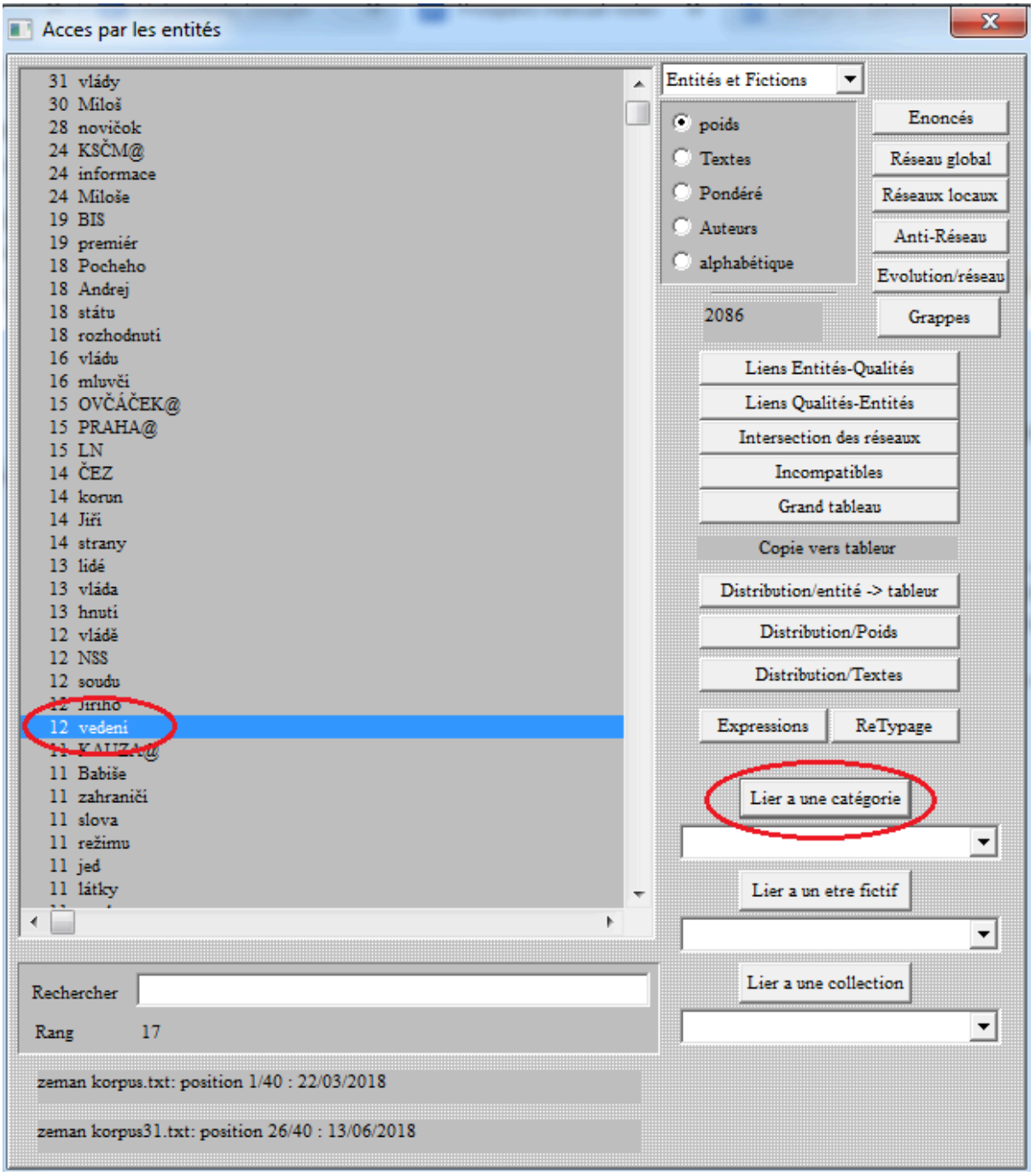
Pokud se nově uložený výraz neobjevuje v příslušném sloupci, je třeba korpus znovu otevřít přes *Gestion du projet – Effacer* a *Ouvrir le projet*.

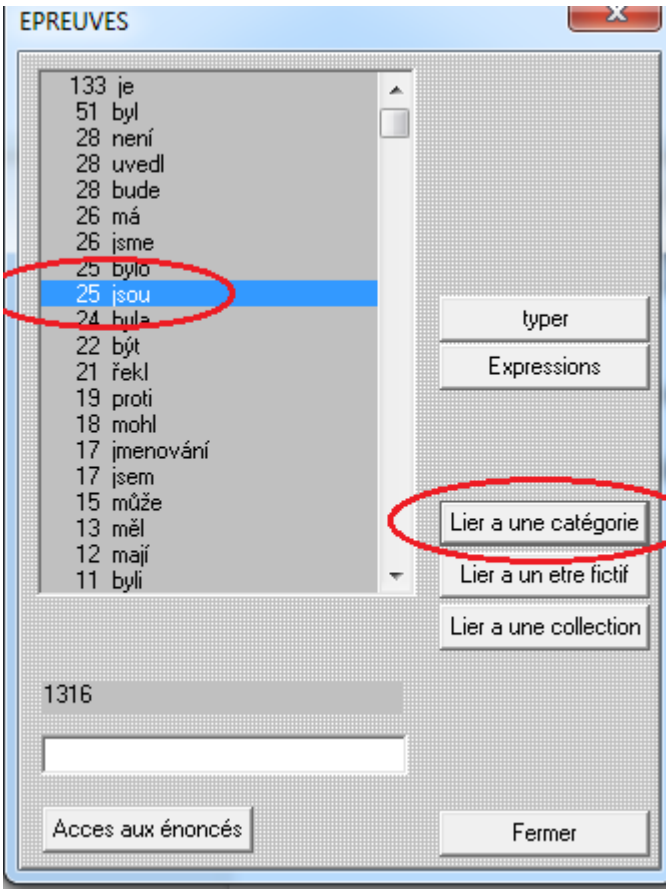
Další možností je uložení kategorie přes *Menu – Outils – Recherche par préfixe/suffixe*. Po vyhledání příslušného výrazu jej označíme, klikneme na pravé tlačítko myši a zvolíme *Lier a une catégorie*. Dále postupujeme jako v minulém případě, pouze s tím rozdílem, že použijeme tlačítko *Lier*:



Slovo takto uložené do kategorie pak není uloženo zároveň do jazykového slovníku jako určitý slovní druh, nýbrž pouze do slovníku pojmového. Prospéro s ním ovšem pracuje, jako by do jazykového slovníku uloženo bylo. Přidáme-li např. slovo nebo slovní spojení k jakékoli kategorii *épreuves*, Prospéro ho bude dále brát jako *épreuve* (sloveso). Ke slovnímu druhu (do jazykových slovníků) je třeba zařazovat slova přes *Modes d'accès – Listes* nebo přes *Typer/Retyper* v jakémkoliv seznamu, viz návod výše.

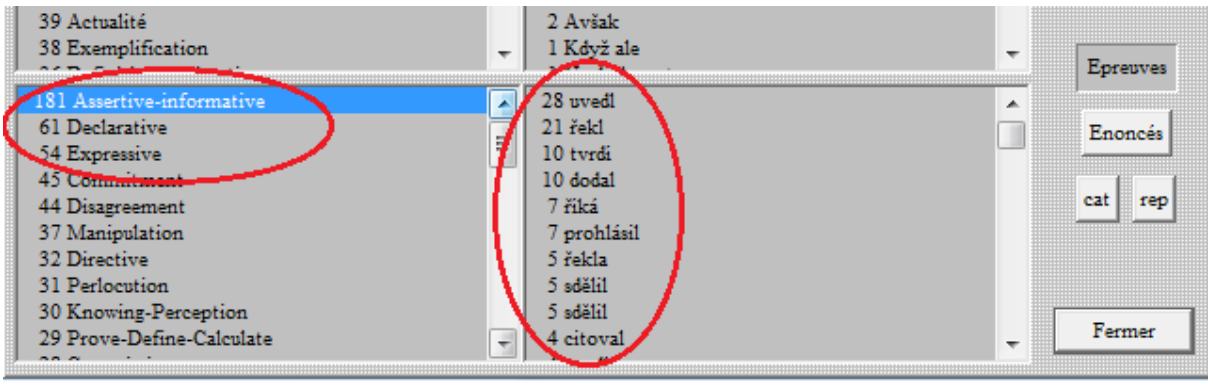
Jazykové i pojmové slovníky jsou flexibilní dle našich potřeb; lze tak přiřadit např. příslovce ***spoustu*** – výraz označující kvantitu – do skupiny *qualités* a dále s ním takto pracovat. Toto se týká jazykových i pojmových slovníků: výrazy nemusí vždy odpovídat slovnímu druhu, jde nám primárně o sémantiku slov, a to i s ohledem na povahu našeho korpusu. Do kategorií lze dále přidávat prvky přes jakýkoliv seznam (*Modes d'accès – Listes*):


Acces par les entités
31 vlády
30 Miloš
28 novičok
24 KSČM@
24 informace
24 Miloše
19 BIS
19 premiér
18 Pocheho
18 Andrej
18 státu
18 rozhodnuti
16 vládu
16 mluvči
15 OVČÁČEK@
15 PRAHA@
15 LN
14 ČEZ
14 korun
14 Jiří
14 strany
13 lidé
13 vláda
13 hnuti
12 vládě
12 NSS
12 soudu
12 Jiriho
12 vedeni
11 Babiše
11 zahraničí
11 slova
11 režimu
11 jed
11 látky
Entités et Fictions
poids
Textes
Pondéré
Auteurs
alphabétique
Enoncés
Réseau global
Réseaux locaux
Anti-Réseau
Evolution/réseau
2086
Grappes
Liens Entités-Qualités
Liens Qualités-Entités
Intersection des réseaux
Incompatibles
Grand tableau
Copie vers tableur
Distribution/entité -> tableur
Distribution/Poids
Distribution/Textes
Expressions
ReTypage
Lier a une catégorie
Lier a un etre fictif
Lier a une collection
Rechercher
Rang 17
zeman korpus.txt: position 1/40 : 22/03/2018
zeman korpus31.txt: position 26/40 : 13/06/2018



Současná podoba kategorií

Pro práci s kategoriemi je klíčové si uvědomit, jakým způsobem byly dosavadní kategorie konstruovány, protože jejich původ představuje určité omezení při jejich užití u odlišných typů korpusu:



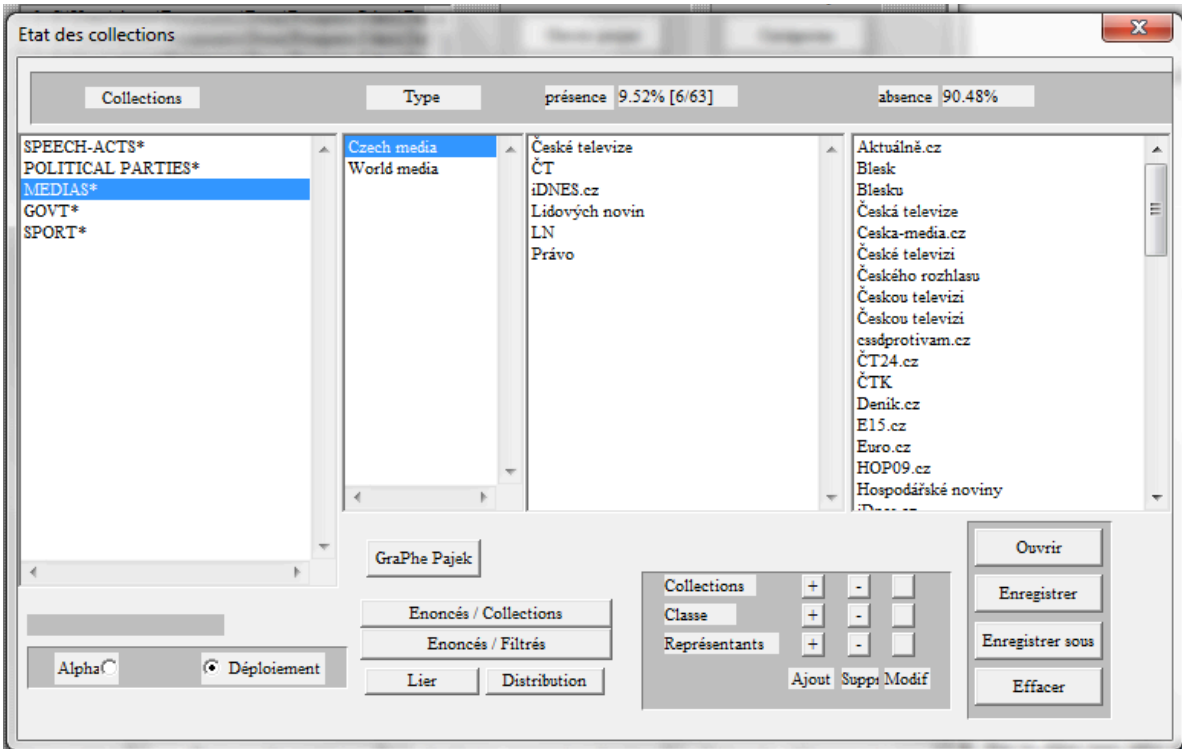
Systém kategorií ve stávající české verzi Prospéra je z velké části výsledkem práce Simona Smithe na třech korpusech především z novinových článků (ale i přepisů rozhlasových pořadů), sestavených v rámci projektu *Narativizace krize* a pocházejících z období od poloviny roku 2009 do parlamentních voleb v květnu 2010: jeden pokrývá

vznik a rozjezd TOP 09, další vznik a rozjezd strany Věci veřejné a třetí zachycuje medializovaný politický diskurz specifického žánru (novinové články vycházející z televizních politických debat).
Proč mají některé kategorie francouzské a další anglické názvy? Francouzské názvy byly zachovány v případech, kde kategorie korespondují s kategoriemi francouzské verze Prospéra. Práce na českých kategoriích byla totiž zčásti inspirovaná dlouhodobou prací Chateauraynaudovy výzkumné skupiny v oblasti sociologie veřejných kontroverzí a sociologie argumentace a považovali jsme za užitečné upozornit na tuto „genetickou stopu“, aby uživatel měl povědomí o tom, odkud daný pojem pochází. Kategorie s anglickými názvy jsou naopak ty, které nemají přímou obdobu v francouzském Prospéru, a vznikly nezávisle v rámci analýzy výše uvedených korpusů. Týká se to v první řadě kategorií zaměřených na narativní spíše než na argumentační analýzu (proto zde najdeme kategorie jako *Manipulation, Commitment, Competence, Performance a Sanction*, korespondující se sémio-narativním schématem Algirdase Greimase, anebo velmi různorodě zastoupenou a frekventovanou kategorii *Actants,* obsahující účastníky a jejich role – aktéra, příjemce, osobu dějem zasaženou apod.). Týká se to také většiny kategorií *épreuves*, kde hlavní (početně největší) kategorie jsou odvozeny od kategorizace řečových aktů podle Johna Searla (například často se vyskytující slovesa jako *uvedl, řekl, tvrdí, dodal* apod. jsou řazena do skupiny *Assertive-informative*).
Uživateli Prospéra samozřejmě nic nebrání, aby si ve své vlastní verzi kategorie založil, upravil, přejmenoval anebo smazal, případně aby si vytvořil několik různých variací pojmových slovníků přizpůsobených jednotlivým korpusům. Právě naopak: je žádoucí adaptovat kategorie vlastním výzkumným otázkám. (Pozn.: nevýhodou přejmenování je, že to ztěžuje sdílení formulí, neboť ve formulích často musíme jmenovat kategorie a další pojmy).
Cenné rady pro práci s kategoriemi poskytuje 17. kapitola Chateauraynaudovy knihy *Prospéro* (ke stažení zde). Příloha manuálu poskytuje stručné definice všech kategorií. V případě nejasností doporučujeme položit dotaz ve fóru uživatelů na prosperologie.org.

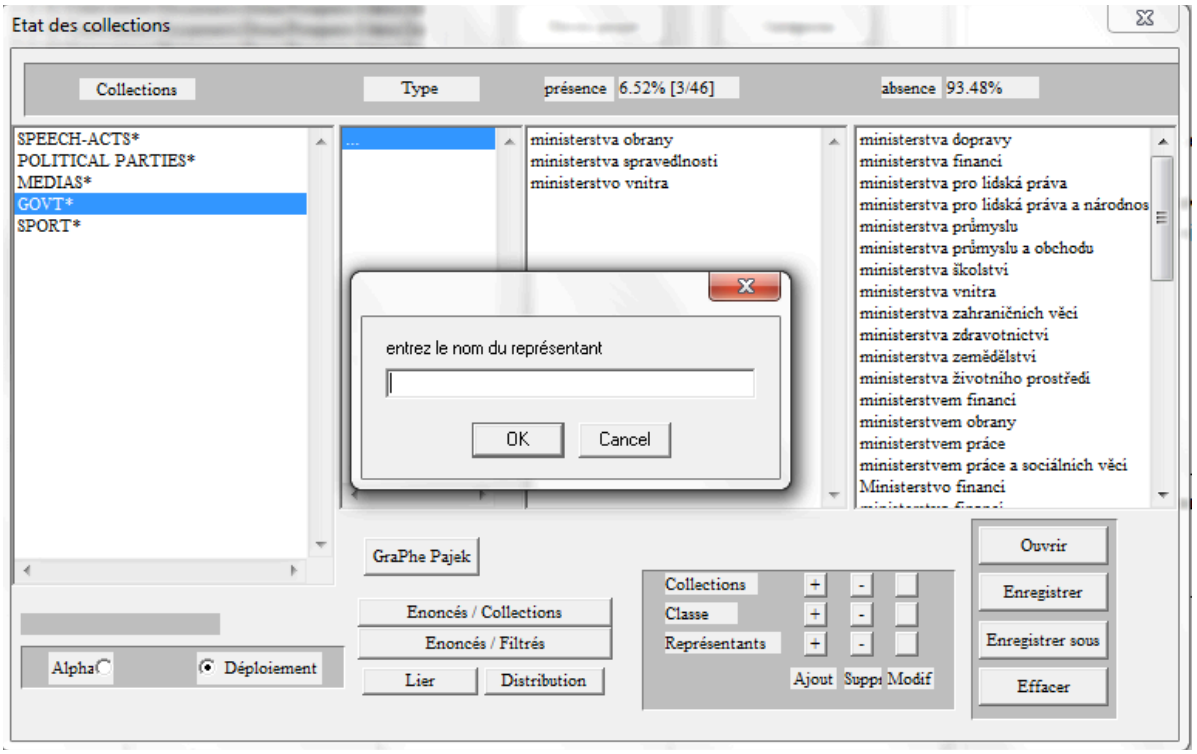
ii. Kolekce

Kolekce zachycují taxonomie, shluky entit po vzoru přírodních věd. Funkce kolekcí se dá nejlépe pochopit jako seznam prvků určitého druhu (země světa, názvy českých médií atp.), jejichž zařazení do dané kolekce je v podstatě nesporné a nezáleží zde na sémantické blízkosti, nýbrž na taxonomickém začlenění. Tematicky zaměřené korpusy občas obsahují částečné seznamy kolekcí, protože autoři-aktéři v rámci svých argumentací mohou vyjmenovat (nejvýznamnější) prvky, např. za účelem uvedení příkladů. Protože relevantnost kolekcí je silně dána zaměřením korpusu, obsahuje instalační balík pouze pár příkladů a je na každém uživateli, aby si doplnil další kolekce relevantní k jeho potřebám. Za tím účelem lze využít encyklopedické zdroje: seznam sportů jsme například čerpali z Wikipedie (ale jen ryze pro ilustraci – aby byla tato kolekce použitelná, bylo by nutné ke každému prvku přidat duplikát s malým začátečním písmenem a rozepsat všechny relevantní pády...).

Prvky kolekcí jsou uvedeny velkými písmeny a opatřeny hvězdičkou:



Opět je lze analogicky přidávat, upravovat apod.:

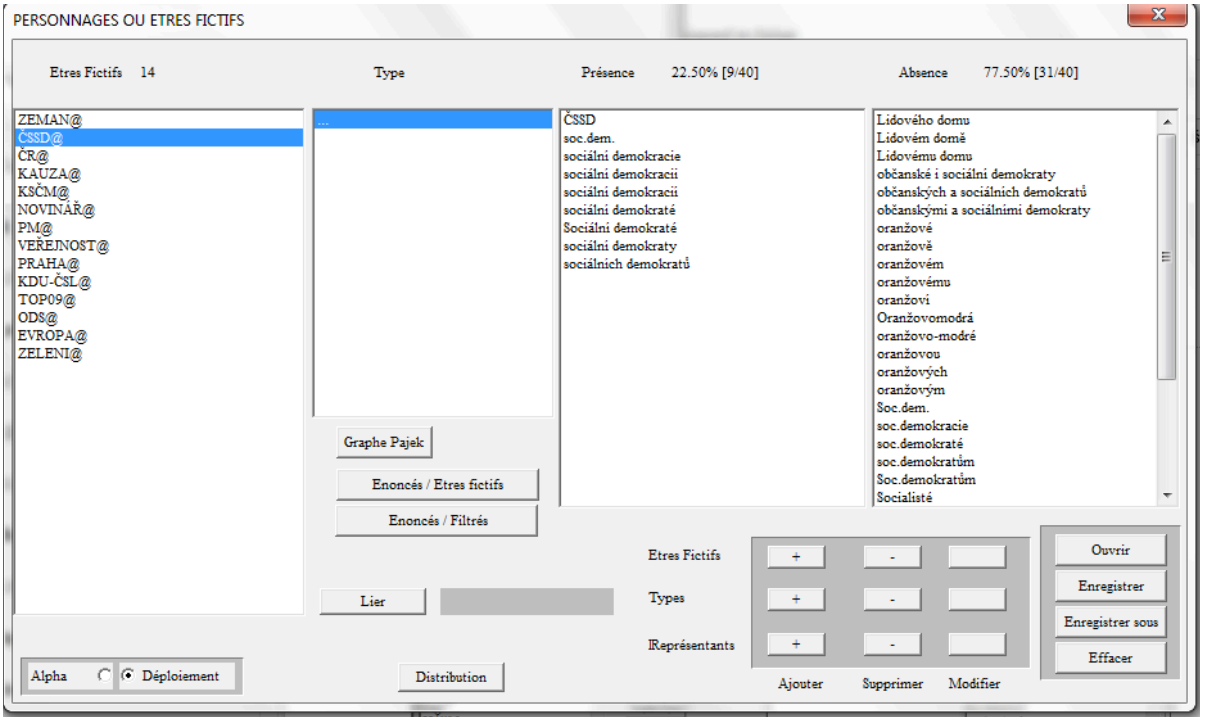


Pozn.: Na rozdíl od kategorií jsou kolekce strukturované trojstupňově: *collection – type/classe – représentant* (kolekce – typ/třída – zástupce). V případě, že nepotřebujeme mezistupeň (jako v případě kolekce GOVT* na obrázku), zadáme v poli *Type/Classe* tři tečky (protože pole nesmí zůstat prázdné).

iii. Velké postavy

Velké postavy korpusu neboli *Etres fictifs (EF)* zachycují praktiky pojmenování nebo označování; jejich hlavním přínosem je sledování jevů jako metonymie, synonymie nebo koreference příznačné pro zkoumaný jazykový svět. Řeší to, že autoři-aktéři často odkazují na stejnou věc různými slovy, ale také umožňují výzkumníkovi sledovat vývoj praktiky jmenování (např. standardizaci nebo vznik nových variací) v čase.

Etres fictifs jsou psány s velkým počátečním písmenem a následovány znakem @. Stejně jako v případě kolekcí poskytuje instalační balík pouze vybrané příklady *EF* relevantní ke cvičnému korpusu. Rozhraní pro práci s nimi je podobné jako u kolekcí a je také strukturované trojstupňově:



Ke všem třem skupinám pojmů lze přistupovat přes menu *Concepts*. Ke kolekcím a kategoriím je možno přistupovat i přes menu *Modes d'accès*. *Etres fictifs* jsou chápány jako podskupina *entités*, proto zde samostatný přístup nemají.

b. Metadata

Metadata jsou strukturované údaje, které nesou informace o primárních datech. Pojem metadat je zde používán především v souvislosti s funkcemi Prospéra. Ten s nimi pracuje

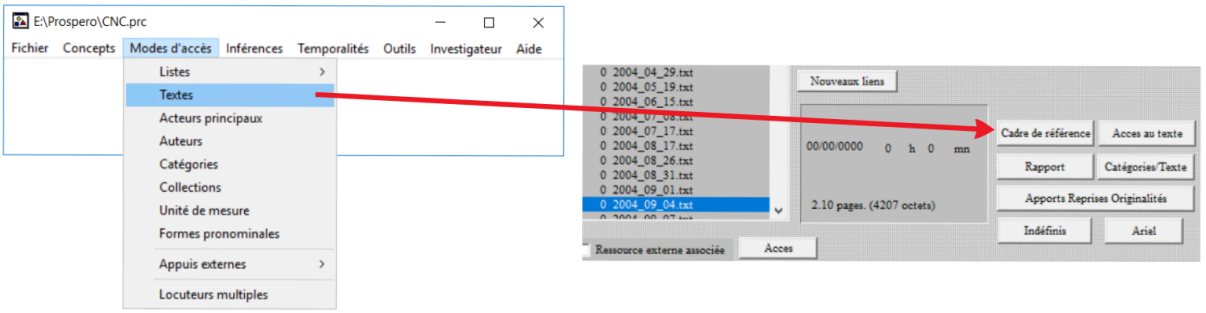
nejen na úrovni popisné, selekční a archivační, ale také v rámci analytických funkcí programu. To znamená, že metadata zde hrají důležitou roli diskriminačních proměnných (viz oddíl Externí/interní korpus a anti-korpus). Metadata Prospéru umožňují rozlišovat výsledky dle autora, jeho statusu (např. politik/novinář), typu/názvu média atd. Tyto informace pak můžeme dále použít, abychom nejen získali představu o článku a korpusu, ale zejména pro porovnání diskursivních jevů vyskytujících se v jiných částech korpusu (v různých časových obdobích, v různých médiích atd.).

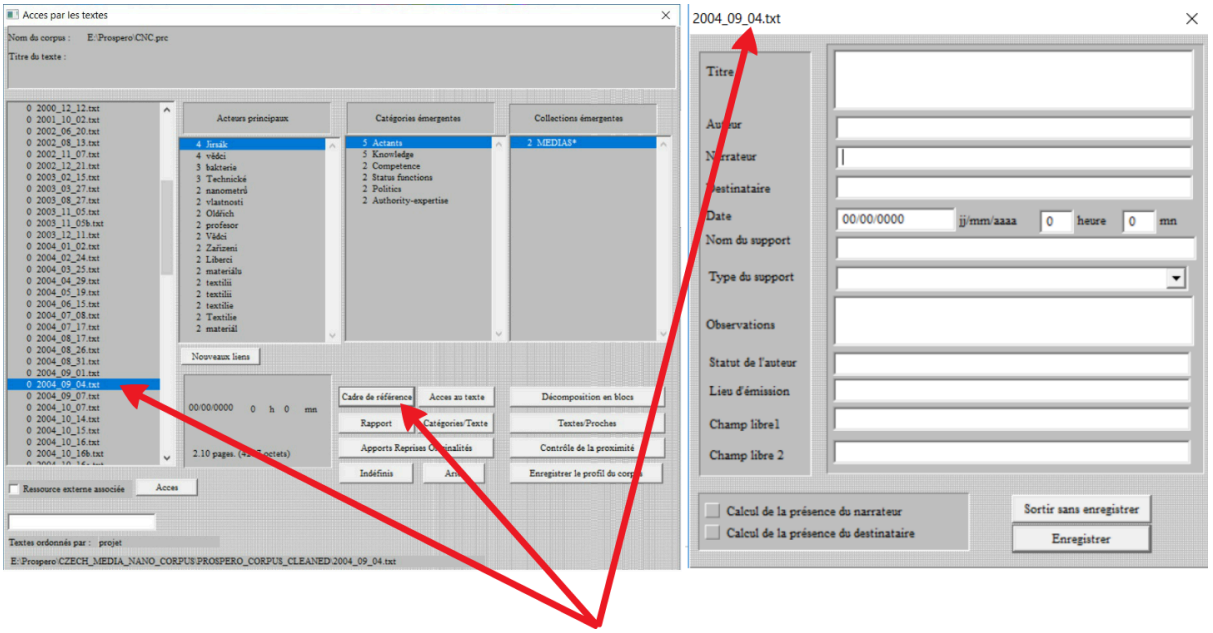
Průvodce tvorbou metadat v prostředí Prospéra

K metadatům se přistupuje dvěma způsoby, které odpovídají dvěma postupům jejich tvorby i zobrazení:

- v uživatelském rozhraní,
- úpravou externího souboru .CTX (typ souboru s metadaty).

Přes uživatelské rozhraní se k metadatům dostaneme postupně přes záložky *Modes d'accès – Textes – Cadre de référence* (viz obrázek). Zde má každý text vlastní záložku metadat (na dalším obrázku vpravo), kde se vyplňují údaje pro vybraný text.



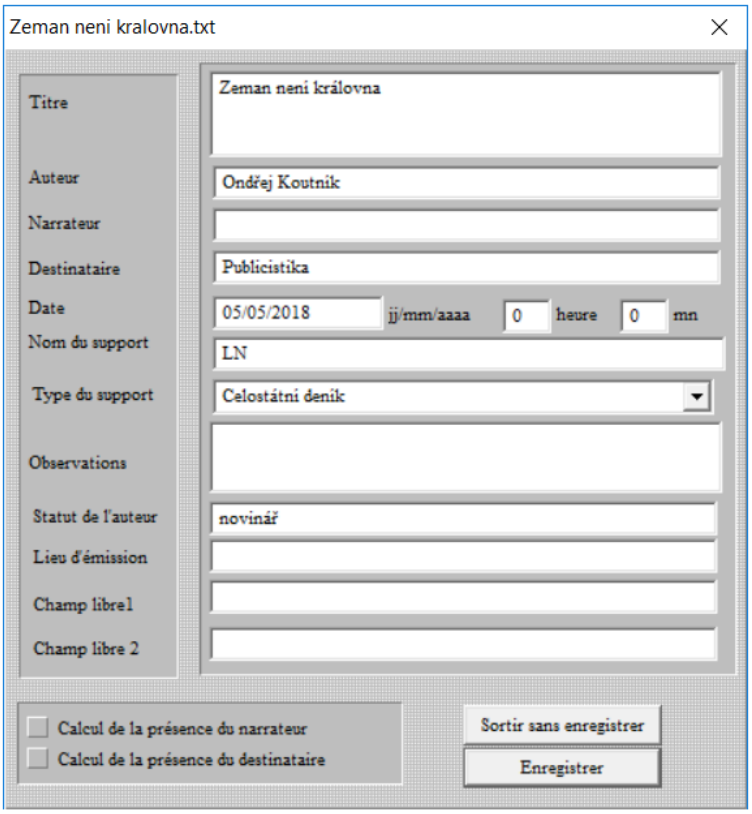


Obrázek ukazuje, jak se pro každý článek korpusu (seznam v levém sloupci uživatelského rozhraní) tlačítkem *Cadre de référence* (uprostřed) otevírá okno položek „metadat“ (vpravo).

V tomto novém dialogovém okně se nachází prázdná pole pro: nadpis (*Titre*), autora *(Auteur*), vypravěče *(Narrateur*), příjemce/adresáta *(Destinataire*), datum (vydání; včetně hodiny a minuty, jedná-li se o audiovizuální zdroj), jméno zdroje (*Nom de support*) (např. Lidové noviny), typ zdroje (*Type de support*) (např. celostátní deník). K dalším údajům patří poznámky *(Observations*), status autora (*Statut de l'auteur*) (např. novinář, politik) a dále místo vydání *(Lieu d'émission*). Poslední dva údaje jsou tzv. volná pole (*Champ libre 1* a *2*).

Obsah těchto polí je plně v režii výzkumníka. Při samotném vyplňování je více než samotná nomenklatura polí důležitá konzistence, neboť jak již bylo uvedeno v úvodu, metadata hrají roli diskriminačních proměnných. Metadata se vyplňují ručně, tzn. zvlášť pro každý text, a i zde se konzistence výzkumníkovi vyplatí.

Například u našeho cvičného textu „Zeman není královna“ se jedná o článek publikovaný v Lidových novinách, 5. května 2004, autorem Ondřejem Koutníkem, formou publicistiky (názorová publicistika). Dalším doprovodným údajem by pak mohla být některá doplňující, ovšem pro výzkum klíčová informace. Například článek mohl vyjít v době po klíčové události, která mohla obsah ovlivnit apod. U tohoto příkladu by údaje mohly být vyplněny následujícím způsobem:



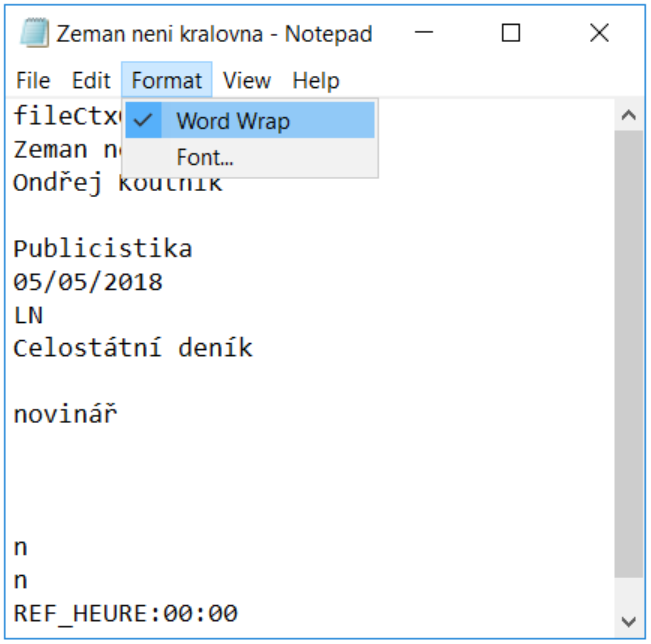
Tlačítkem *Enregistrer* (uložit) se okno zavře a veškeré informace se uloží do nového souboru s názvem „zeman neni kralovna.CTX“.

Doporučení: Při vyplňování větších korpusů se vyplatí průběžně ukládat celý projekt (přes záložky *Fichier – Gestion du projet – Enregistrer projet*). Soubory .CTX jsou vygenerovány po každém „uložení“ metadat u každého vytvořeného záznamu (resp. uloženého, na obrázku *enregistrer*). Umístěny jsou vždy v adresáři se soubory článků.

Poznámka: Pokud jsme při vyplňování metadat udělali náhodnou nebo dokonce systematickou chybu (např. zaměnili jsme jméno autora; chceme doplnit/změnit označení „neznámého autora“, název/typ média apod.), máme možnost opravit záznamy pro vybrané skupiny textů přes záložku *Outils – Recodage des références externes.*

Druhou metodou je **tvorba a úprava metadat přímo v souboru .CTX**. Zde máme opět dvě možnosti:

- Soubor můžeme nejprve vytvořit sami v textovém editoru ve formátu .txt a následně přejmenovat na .ctx.
- Můžeme přepisovat obsah u již existujícího .ctx souboru a průběžně ukládat pod jménem příslušného .txt (nebo nového .ctx) souboru.



Obrázek ukazuje soubor „Zeman neni kralovna.CTX“ otevřený v poznámkovém bloku (v textovém editoru typu Notepad). Každý .CTX soubor obsahuje záhlaví „fileCtx0005“ a na konci textu dva znaky „n“ na samostatných řádcích. Na konci by měly být dvě prázdné mezery – záleží na tom, zdali záznam obsahuje časový údaj, uváděný v uživatelském rozhraní v poli *date*.

Důležité upozornění: Při přímé editaci souboru .CTX je třeba dbát na zachování počtu řádků a ověřit tzv. zalamování řádků (Word Wrap) přímo pod záložkou *Formát* (také na obrázku výše). V nezalomeném textu by totiž nebyly rozlišitelné jednotlivé záznamy metadat.

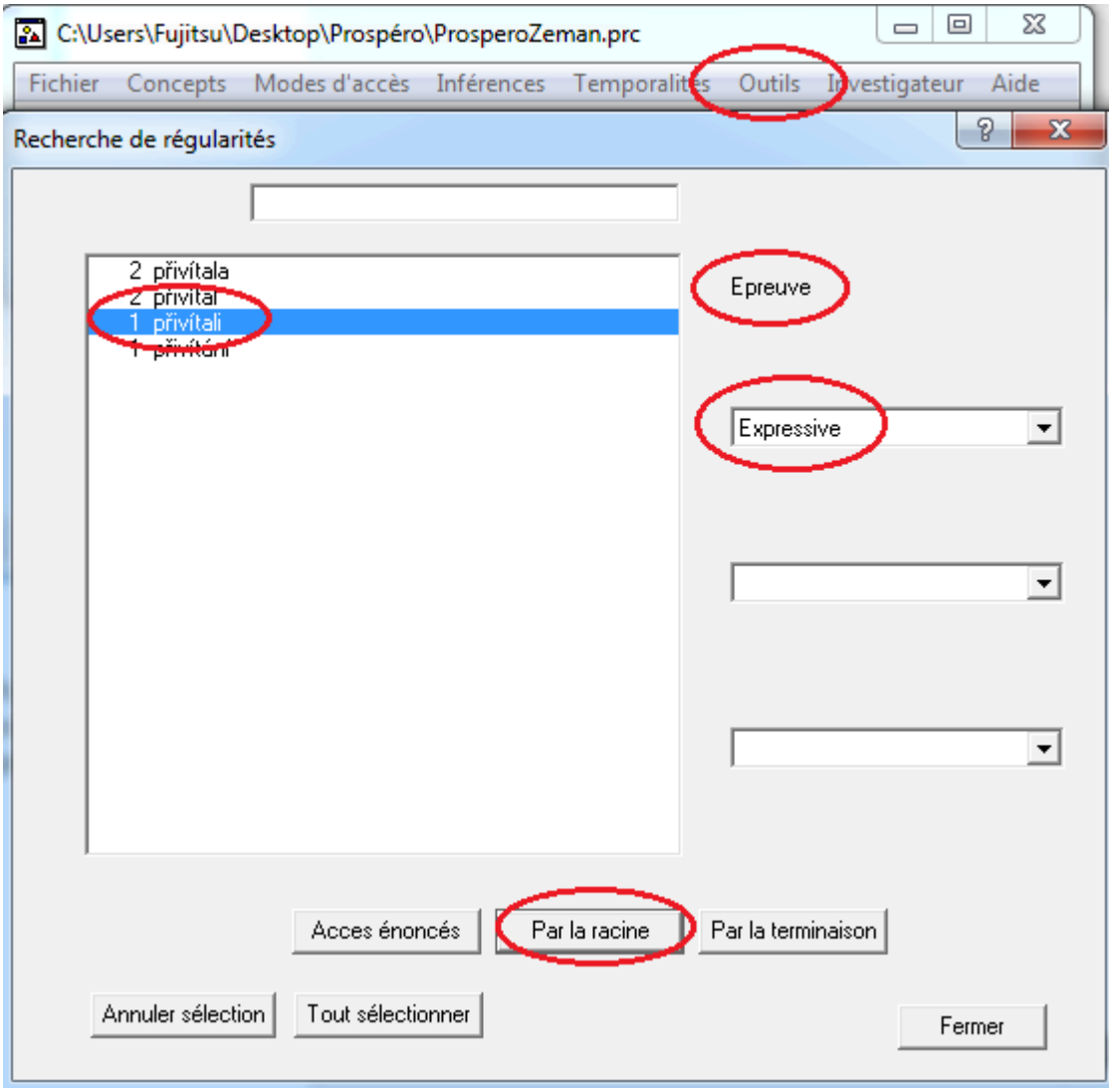
Vytvořenému/upravenému souboru přiřadíme stejný název, jaký má příslušný soubor s textem článku. Oba soubory musí být umístěny ve stejné složce se zdrojovými texty korpusu (také na obrázku níže). Prospéro na základě této shody dokáže propojit obsahy textů (.TXT) s metadaty (.CTX).

c. Analytické nástroje Prospéra

Prospéro nabízí několik analytických nástrojů, z nichž v tomto oddílu probereme několik nejužívanějších.

i. Vyhledávač příbuzných výrazů

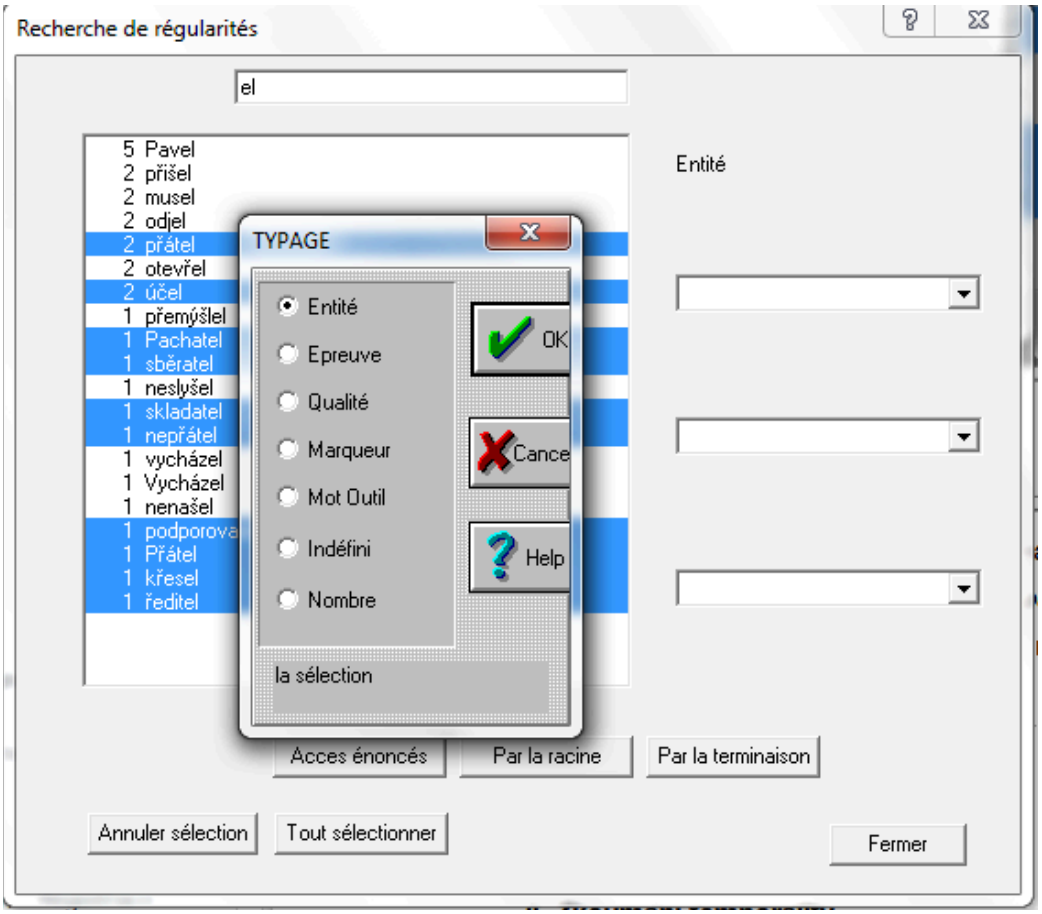
Užitečnou funkcí Prospéra pro orientaci v korpusu je vyhledavač libovolného slova a jeho odvozenin či slov příbuzných, který nalezneme v hlavním menu pod označením *Outils – Recherche par préfixe/suffixe*:



Do vyhledávacího okna zadáme slovo či jeho část a vybereme mezi dvěma parametry vyhledávání: *par la racine* zde znamená dle začátku slova; *par la terminaison* pak podle jeho koncovky. V našem příkladě byla hledána část slova ***přiví*** jako *racine* a byly nalezeny čtyři související výrazy. Např. sloveso ***přivítali*** se pak v našem korpusu vyskytuje jedenkrát (číslo před slovem) a je klasifikováno jako *épreuve* (slovní druh) a *expressive* (kategorie; viz níže). Za pomoci této funkce tedy můžeme zjistit, do jaké kategorie je daný výraz zařazen. Dále se opět můžeme pomocí *Acces énoncés* (spolu s označením vybraného nalezeného výrazu kurzorem) podívat přímo na kontext zvoleného slova.

Důležité také je, že zde můžeme označit více výrazů najednou, a takto hromadně je tedy můžeme např. také klasifikovat do zvolené množiny slov. V kapitole „Jak Prospéro čte vložené texty?“ jsme ukázali manuální indexaci slov po jednom, přičemž zde se nabízí postup jak to udělat rychleji. V češtině je to obzvlášť vhodný způsob pro téže slovo v různém pádu, lišící se pouze v rodě apod. V našem příkladě bychom všechna čtyři slova pravděpodobně zařadili k *épreuves*.

Indexace slov pomocí vyhledavače se dobře uplatňuje taky pro některé sporné české koncovky. Například dictio.cfg je nastavený tak, aby automaticky zařadil slova končící *-el* k *épreuves*. Tím pádem špatně klasifikuje četná podstatná jména (zejména ta s koncovkou -tel). Hodí se tak pro každý nový korpus vyhledat slova končící *-el* a hromadně překlasifikovat podstatná jména ze seznamu jako *entités*. Příklad je z našeho korpusu:



Pozn. Jména jako *Pavel* nemusíme klasifikovat: pokud jsou v souboru prenoms.dic, budou automaticky zařazena k entités, a pokud tam nejsou, je třeba je tam přidat!

Postupem času, jak obohacujeme české slovníky na základě co nejheterogennějších

korpusů, bude potřeba vykonat tuto operaci klesat, hledání *par la terminaison* je nicméně každopádně dobrou kontrolou. Je doporučováno to provést aspoň pro následující sporné koncovky (tučně jsou označeny klasifikace podle souboru dictio.cfg):

-á (**qualité** / épreuve)

-et (**épreuve** / entité nebo nombre)

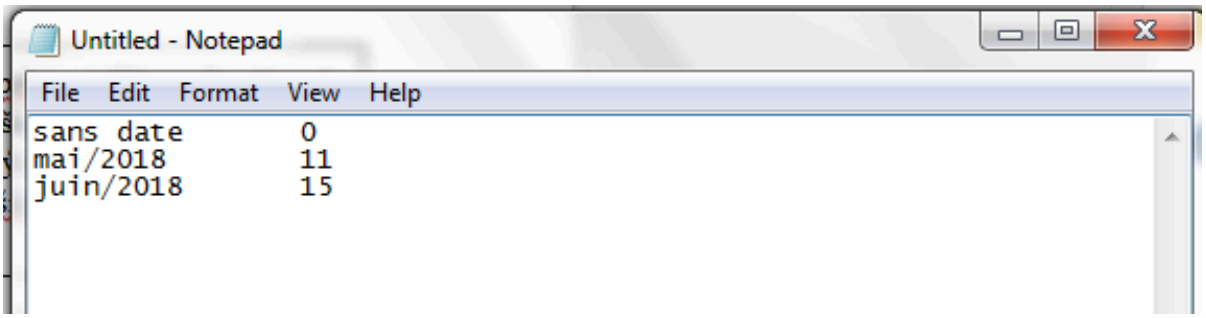
-el (**épreuve** / entité)

-ně (**marqueur** / entité)

ii. Zkoumání temporality

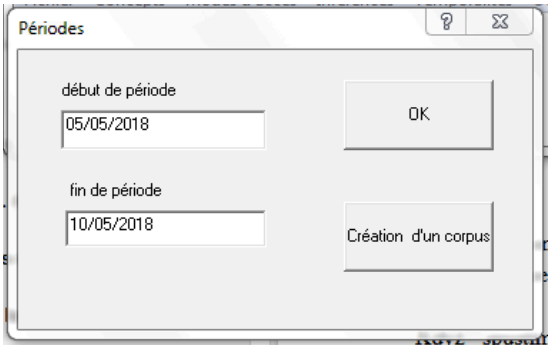
Čas je ústředním motivem pragmatické sociologické školy, z níž Prospéro vzniklo, a mnoho z kategorií zejména *marqueurs* bylo vytvořeno za účelem zkoumání způsobů, jak autoři textů našich korpusů „hospodaří“ s časem diskurzivně – Jak se orientují vůči budoucnosti, minulosti a přítomnosti, jak mobilizu jí precedenty a prognózují anebo jak modalizují temporalitu dění tím, že naléhají („času je málo“) anebo naopak vyčkávají („času je dost“). Kromě těchto a dalších diskurzivních znaků nabízí Prospéro sadu funkcí, jejichž prostřednictvím můžeme získat různé statistické nebo kvazi-statistické míry časového rozvržení korpusů. Za předpokladu, že máme pro každý text vyplněné nebo vygenerované minimálně jedno metadatum – a to datum jeho vzniku nebo vydání. V tomto oddílu nabízíme krátký přehled nástrojů temporality. Většinou se nacházejí pod záložkou *Temporalités.*

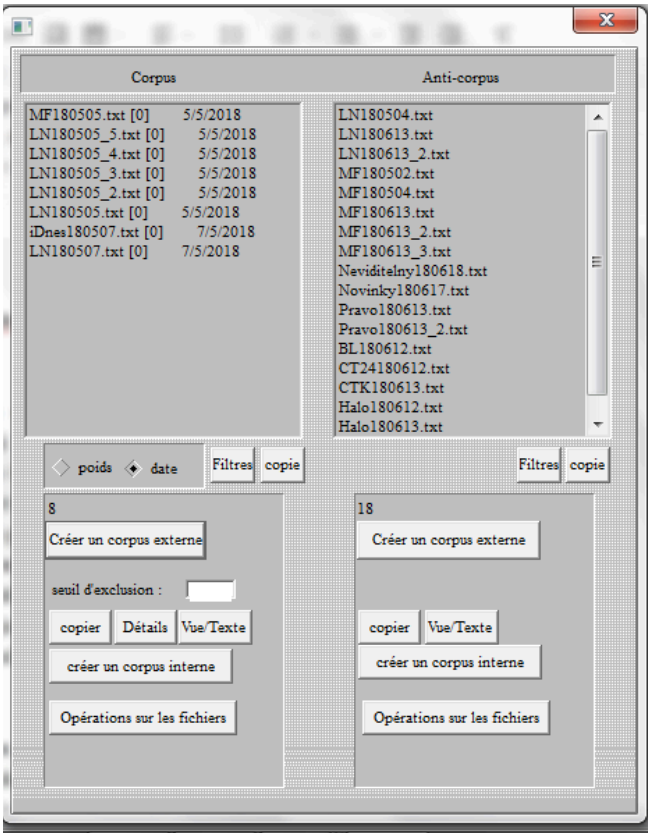
Spuštěním příkazu *Temporalités – Temps du corpus* můžeme získat jednoduchý přehled rozvržení korpusu podle dní, měsíců, roků nebo libovolného násobku těchto jednotek. Prospéro nám zkopíruje tabulku ukazující počet textů za vybraný interval, kterou vložíme do textového editoru, případně do Excelu. Názvy měsíců jsou ve francouzštině.



Obdobnou tabulku lze vytvořit pro distribuci základních pojmů (*entités* – včetně *Etres-fictifs*, *qualités*, *marqueurs*, *épreuves*), pokud na daný pojem v seznamu *entités* klikneme pravým tlačítkem myši a zvolíme *copier élément/temps*. Stejný výsledek pro kategorie *entités* obdržíme po spuštění příkazu *Modes d'accès – Catégories* kliknutím na jednu z kategorií a pak na tlačítko *Sortie vers tableur* pod hlavičkou *Distribution de la catégorie*.

Příkaz *Temporalités – Périodisation manuelle* otevře dialogové okno, kde nastavujeme začátek a/nebo konec období, pro něž chceme vytvořit podkorpus. Pokud máme důvod se domnívat, že se určitý časový úsek našeho korpusu může v něčem lišit, můžeme porovnávat jeho diskurzivní rysy s dalším obdobím nebo s celým korpusem. Takto lze například zkoumat, jestli po nějaké klíčové události došlo k zásadní změně v jazyce aktérů nebo v jejich způsobu argumentace.

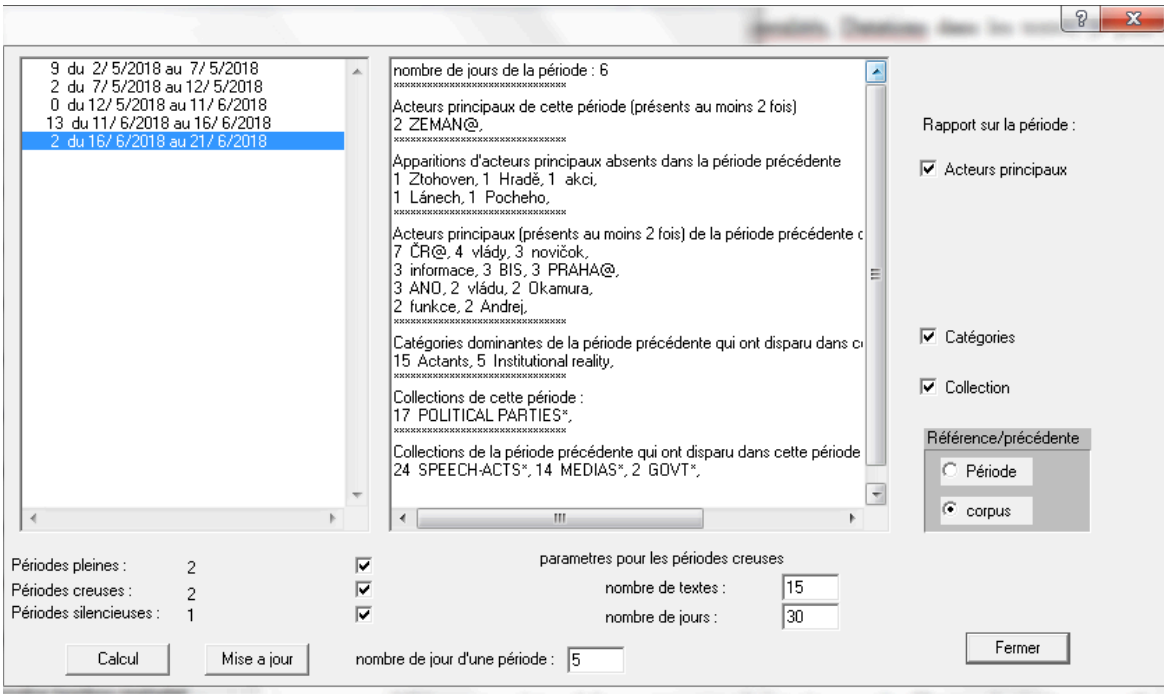




Třetí funkci pod záložkou *Temporalités, Datations dans les textes* necháme stranou, neboť je bohužel plně funkční pouze pro francouzský jazyk.

Když spustíme příkaz *Temporalités – Calculs des périodes,* rozdělíme korpus do rovnoměrných úseků, tentokrát podle počtu dní (*nombre de jour d'une période* – přednastaveným parametrem je hodnota 20). Kliknutím na tlačítko *Calcul* (viz obrázek níže)

pak obdržíme v levém sloupci seznam období s počty textů. V obrázku vidíme odpočet tzv. plných (*pleines*), dutých (*creuses*) a tichých (*silencieuses*) období podle námi vybraných parametrů. Na pravé straně okna můžeme zadat zprávy za jednotlivá období a zjistit, které byly hlavní postavy (*acteurs principaux*), dominantní kategorie a zastoupené kolekce, a to buď v porovnání s celým korpusem, anebo s předchozím obdobím.

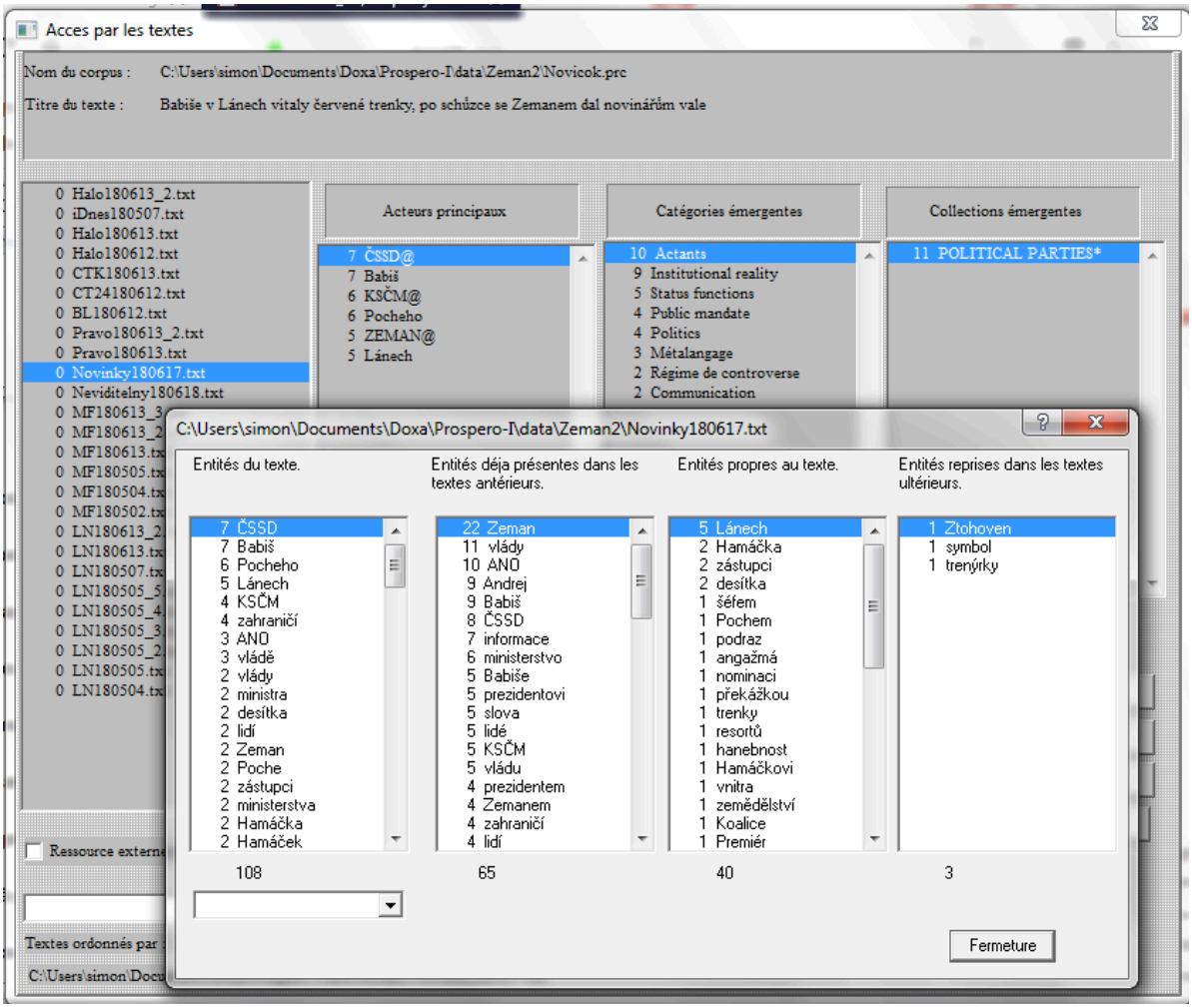


Prospéro definuje tiché období jako období bez žádných textů. Hranice mezi *pleines* a *creuses* je přednastavená na 5 textů za 30 dní. To je pochopitelně vhodné přizpůsobit časovým rozměrům každého korpusu – v příkladu jsme si hranici zvýšili na 15 textů za 30 dní.

Další časová funkce, *Temporalités – Episodes,* představuje poněkud hravý nástroj, s jehož pomocí Prospéro vygeneruje předem nespecifikovaný počet období podle tří různých algoritmů (je třeba nejprve vybrat *Calcul* A, B nebo C a pak kliknout na *Résultats*). Každý algoritmus má za cíl identifikovat přelomové texty podle změn hlavních postav. Podrobné zprávy, které obdržíme vpravo kliknutím na tlačítko *Détails*, se týkají vždy prvního textu v daném období, který může být zajímavý právě tím, že nastolil změnu hlavních postav, a z citací jsou vybrány ty, v nichž figurují hlavní postavy období. Obvykle se jedná o velmi krátká období.

Podobnou operaci můžeme provést i v obráceném smyslu. Chceme-li zjistit, co nového přinese konkrétní text, přes panel rozhraní *Modes d'accès – Textes* vybereme jeden

text a klikneme na tlačítko *Apports Reprises Originalités*. Potom vybereme jeden druh základních pojmů z nabídky v dolní levé části okna (např. *entités*) a uvidíme (zleva doprava): *entités* vyskytující se v textu; *entités* vyskytující se v textu již zmíněné v jednom nebo více dřívějších textů; *entités* vyskytující se výhradně v tomto textu; a *entités* vyskytující se poprvé v tomto textu a ještě alespoň v jednom pozdějším textu. Zajímavost textů může totiž spočívat v tom, že zavedly do korpusu nové pojmy, anebo naopak že synteticky shrnuly dosavadní debaty a pojmové repertoáry. Na obrázku níže například vidíme, že slova *Ztohoven* a *trenýrky* jsou zmíněna poprvé ve zprávě vydané na serveru novinky.cz a následně se vyskytla i v dalších textech.



Poslední nástroj na časovou analýzu korpusu umožňuje sledování postupných změn v konfiguraci vztahů kolem vybrané *entité* (nejlépe kolem *être-fictif*). Ze seznamu *entité*, znovu přes záložku *Modes d'accès*, vybereme některou položku a klikneme na tlačítko *Evolution/réseau*, pak na *Reconfigurations successives*. Prospéro automaticky rozděluje korpus do několika období – zakládá nové období tehdy, když dojde k 75procentní změně v kompozici sítě silně propojených *entités* (na obrázku v pravém sloupci).

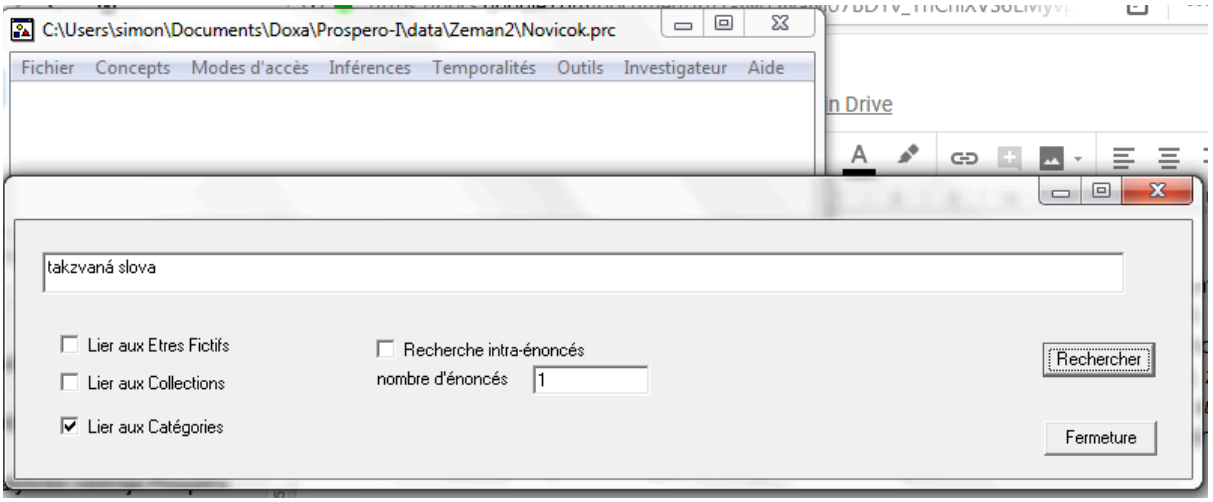
iii. Konfigurace pojmů

Existují dva nástroje ke zkoumání vztahů mezi kategoriemi, kolekcemi a velkými postavami, které nám umožňují identifikovat text anebo pasáž, kde se vyskytují vybrané konfigurace pojmů - *explorateur interne* a *configurations discursives*.

Explorateur interne

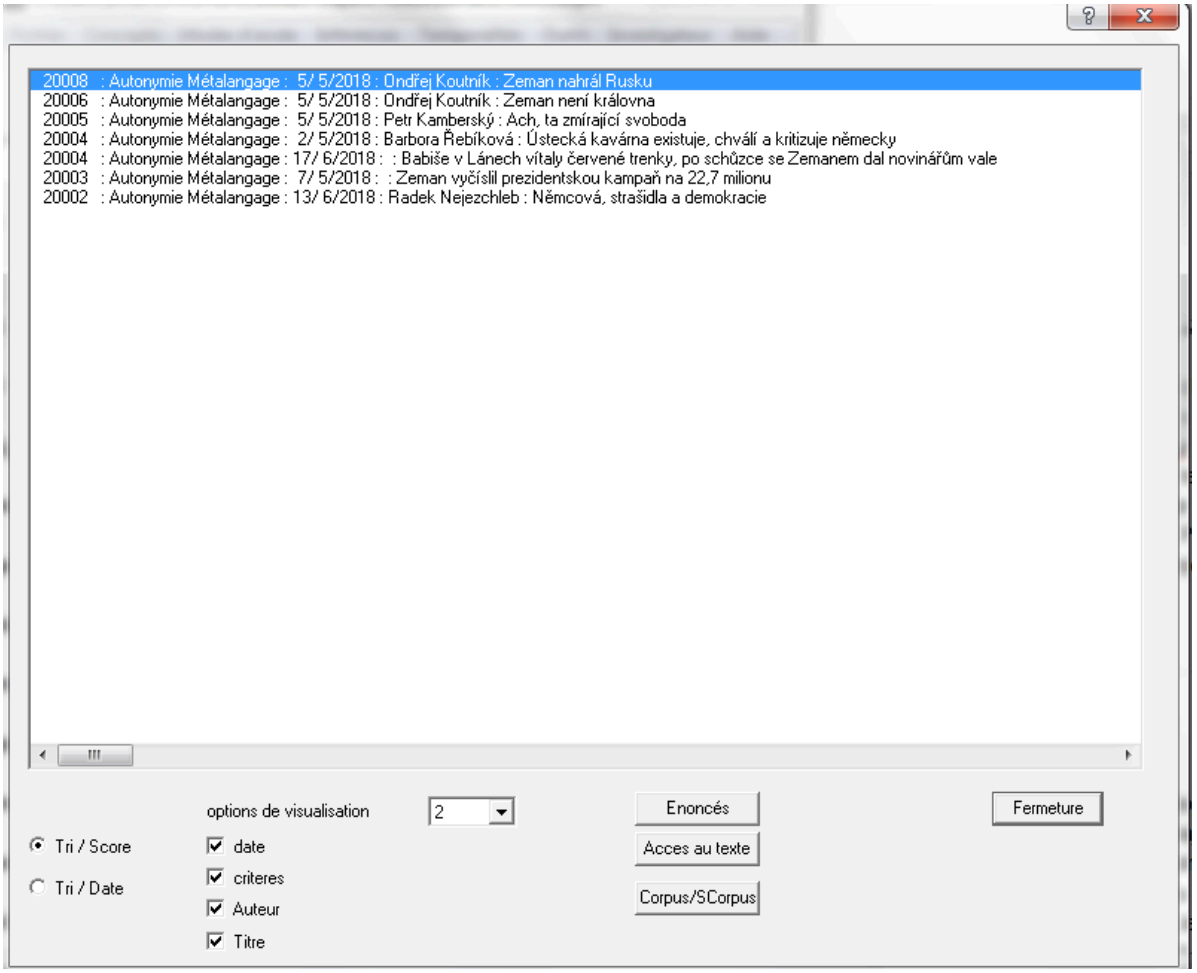
Spustíme-li příkaz *Outils – Explorateur interne*, otevře se okno s dlouhým příkazovým řádkem, do nějž můžeme vložit dvě nebo více slov (oddělených mezerami) nacházejících se v korpusu. Prospéro pro nás pak hledá kolokace mezi těmito slovy. Máme ovšem možnost proces vyhledávání posílit podle toho, k jakým konceptům tato slova patří. To znamená, že po zaškrtnutí jednoho (nebo více) z políček *Lier aux Etres fictifs*, *Lier aux Collections* a *Lier aux Catégories* bude Prospéro hledat kolokace nejen mezi danými slovy, ale také mezi jakýmikoli zástupci konceptů vybraného typu, ke kterým daná slova patří.

Dejme tomu, že se zajímáme o metajazykové výpovědi aktérů v korpusu. Mohli bychom vybrat dva zástupce kategorie *Métalangage* a *Autonymie*, např. ***slova*** a ***takzvaná***. Zadáme tedy do příkazového řádku ***takzvaná slova***, zaškrtneme políčko *Lier aux catégories* a klikneme na *Rechercher*.

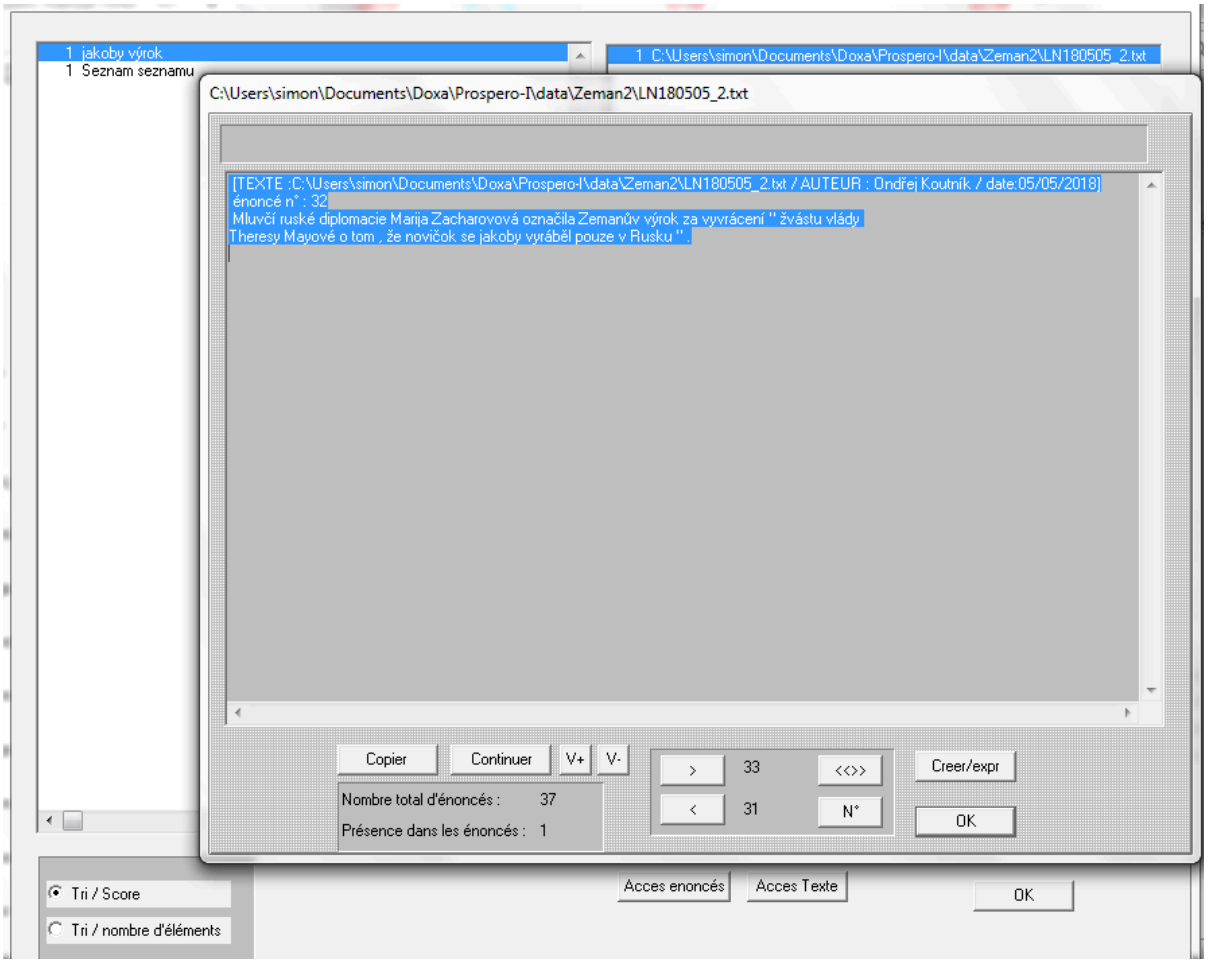


Potom (viz obrázek níže) je možné upřesnit výsledky tak, abychom viděli jen texty, ve kterých se vyskytují obě kategorie zároveň (vybrat *options de visualisation* – 2), a abychom viděli autory a názvy článků (zaškrtnout políčka *Auteur* a *Titre*). Dále máme možnost označit buď *Tri / Score,* anebo na *Tri / Date* podle toho, jestli chceme uspořádat výsledky podle počtů výskytů v textech, nebo chronologicky. Čísla na levé straně naznačují, kolik kategorií a kolik výskytů je v jednotlivých textech (např. kód 20008 znamená 2 kategorie a

dohromady 8 výskytů). Pro zobrazení jednotlivých výskytů klikneme na *Enoncés* a navigujeme v nich podle už známých pokynů.



Explorateur interne nám pak umožňuje zpřísnit parametry tak, aby se daná slova nebo dané pojmy vyskytly nejen ve stejných textech, ale i ve stejných větách nebo skupinách vět. K tomu zavřeme aktuální okno a v předchozí nabídce označíme *Recherche intra-énoncés*. Necháme-li vedle *nombre d'énoncés* zadané číslo 1, provede Prospéro nejnáročnější test kolokace – slova/pojmy musí být ve stejné větě. V daném případě jde o dvě věty v korpusu, z nichž pouze ta první představuje metajazyk v pravém slova smyslu (výpovědi o významu slov).



Configurations discursives

Druhý nástroj na zkoumání vztahů mezi pojmy operuje výhradně na úrovni textů. Je vhodnější než *Explorateur interne,* pokud máme přesnou představu o tom, jaká konfigurace nás zajímá, a obzvlášť pokud chceme různě vážit jednotlivé pojmy uvnitř konfigurace, což *Explorateur interne* neumožňuje. Jde o vymezenou sadu nástrojů pod záložkou *Inférences: Régimes Discursifs* (pro konfigurace kategorií), *Jeu d'acteurs* (pro konfigurace velkých postav), *Répertoires* (pro konfigurace kolekcí) a *Configurations discursives* (pro konfigurace sestavené z více druhů pojmů). Všechny fungují a vypadají podobně.

Po spuštění vybraného nástroje je třeba kliknout na tlačítko *Créer* a pojmenovat konfiguraci. Potom je třeba tuto konfiguraci editovat (tlačítko *Editer*).

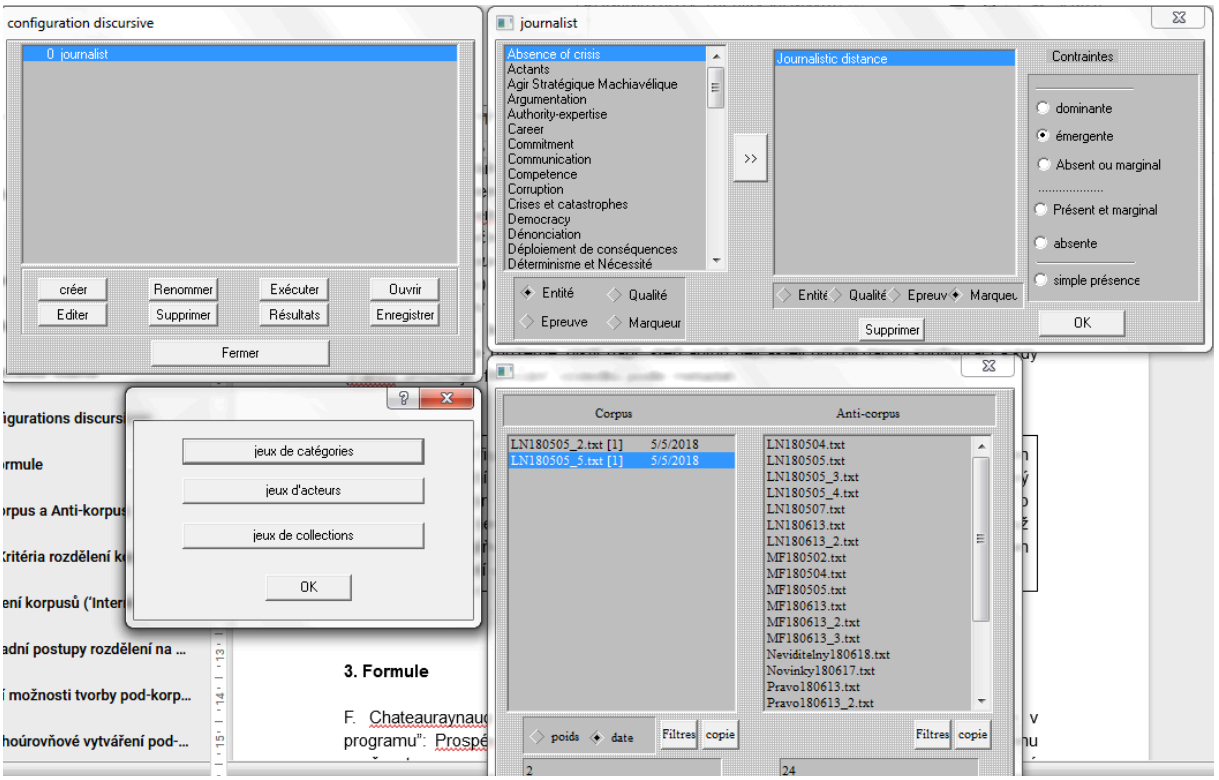
Pozn.: Pouze pro *Configurations discursives* se pak objeví dialogové okno, v němž upřesníme, z čeho chceme vybrat: z kategorií, kolekcí nebo velkých postav

Následně vybereme ze seznamu v levém okně minimálně dva pojmy, které budou součástí konfigurace, a kliknutím na dvojitou šipku je přesuneme do pravého okna.

Poslední krok v nastavení konfigurace je vážení pojmů, kde máme pět voleb, které jsou (od největší váhy k nejmenší): *dominante, émergente, présente, marginale* a *absente*. Často je nejvhodnější orientovat se podle počtu výsledků, který obdržíme, a pak upravovat nastavení konfigurace, dokud není výběr zhruba tak náročný, jak potřebujeme. Někdy například hledáme pár klíčových textů, kde je daná konfigurace nejsilněji zastoupená, a jindy chceme identifikovat všechny texty, kde je jen přítomná.

Poté, co jsme vybrali naše pojmy, potvrdíme tlačítkem OK a vrátíme se tak k předchozímu rozhraní. Zde je třeba kliknout na *Exécuter* a konečně na *Résultats*. Výsledky jsou prezentovány jako dva seznamy označené *Corpus* (všechny texty vyhovující našim kritériím) a *Anti-corpus* (všechny nevyhovující texty). Manipulace s podkorpusy je popsána podrobně v následujícím oddílu iv, zde si vystačíme s jednoduchým zobrazením výsledků. Jednotlivě si texty můžeme prohlédnout po kliknutí na *Vue/Texte* (*Détails* zde funkční, protože Prospéro prohledal celé texty, nikoli věty). Kliknutím na tlačítko *Filtres* dostaneme další užitečný přehled, kde můžeme zjistit např. to, kteří autoři nejčastěji použili danou konfiguraci a kdy (*Filtres* umožňuje filtrování výsledků podle metadat).

Pro uchování našich konfigurací je musíme uložit (tlačítko *Enregistrer*).

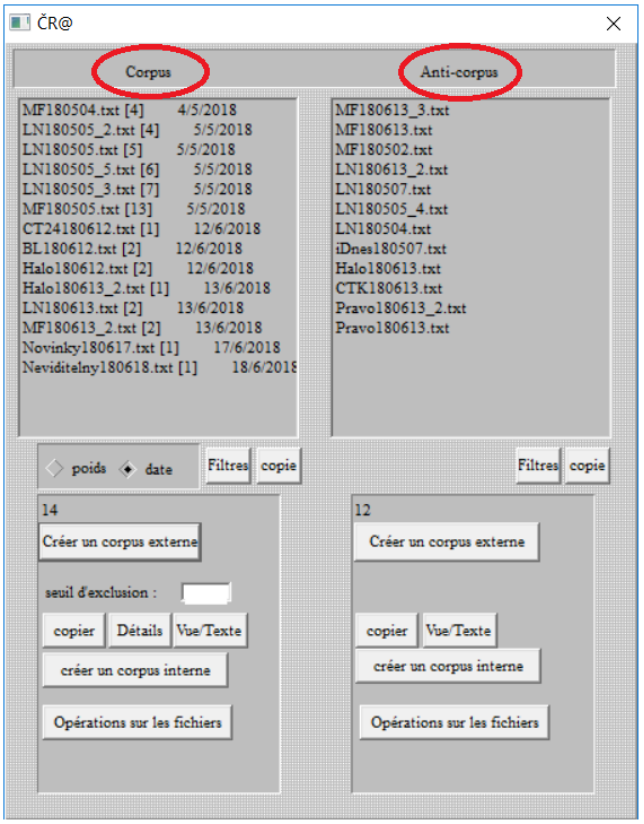


Tip: Pomocí nástrojů konfigurace pojmů můžeme také zjišťovat, ve kterých textech je určitý pojem dominantní nebo emergentní. K tomu vytvoříme konfiguraci obsahující pojem, který nás zajímá, vážený jako *dominante* (případně *émergente)*, a k němu přidáme další pojem, o němž víme, že není zastoupen v korpusu. Tento druhý pojem vážíme v konfiguraci jako *absente*. Jelikož všechny texty z definice splňují kritérium nepřítomnosti druhého pojmu, pak bude výsledný seznam obsahovat jen ty texty, v nichž je první pojem dominantní/emergentní.

iv. Dílčí korpus a antikorpus

Jednou z dalších analytických funkcí Prospéra je možnost **rozdělit** a **porovnávat** skupiny textů z hlediska jejich obsahů, např. podstatných jmen (*entité*), vlastností (*qualité*) a modalit (*marqueur*), ale i na základě analytických konceptů, tj. kategorií (*catégories*), kolekcí (*collections*) a velkých postav (*être fictifs*), nebo na základě metadat (autorství, čas, zdroj, žánr apod.)

Tato fáze práce s Prospérem je velmi důležitá, stejně tak jako je běžná v kterémkoli momentu výzkumu/bádání. Během práce s programem má totiž výzkumník kdykoli možnost vrátit se zpět do textu a zkontrolovat užití slova či výrazu, aby určil podmínky jeho výskytu nebo ověřil hypotézu.



Dílčí korpus a antikorpus jsou dva seznamy textů (původního korpusu) rozdělené dle kritéria stanoveného výzkumníkem. Na obrázku je korpus rozdělen dle výskytu @ČR (velké

postavy). Jedná se zde o texty, kde se vyskytují známé slovní tvary pro ***Českou republiku***, např. ***Česká republika***, ***České republice,*** ale i ***Česko*** apod., tvoří tzv. dílčí korpus (na obrázku sloupec vlevo). Tento soubor textů je oddělen od zbytku velkého korpusu, kde se dané výrazy nevyskytují, tzv. antikorpusu (na obrázku sloupec vpravo).

Kritéria a možnosti rozdělení korpusu

Kritéria pro rozdělení základního korpusu na porovnávané „podkorpusy" dílčí mohou být různá, ale vždy zahrnují přítomnost (resp. absenci) určitého slova, slovního spojení, kategorie, sbírky apod. Jinými slovy, dílčí korpus (anti-korpus):

- zachycuje a porovnává: kde, kdy, a jak něco figuruje, resp. nefiguruje,
- ukazuje dílčí (i obecnou) orientaci korpusu z hlediska zvoleného kritéria (o tom později v další části této podkapitoly),
- může být vytvářen na úrovni metadat, tzn. z hlediska přítomnosti autorů, data vydání, místa vydání, typu média apod.,
- je jednou ze základních orientací v korpusu, která slouží k další interpretaci dokumentů (textů).

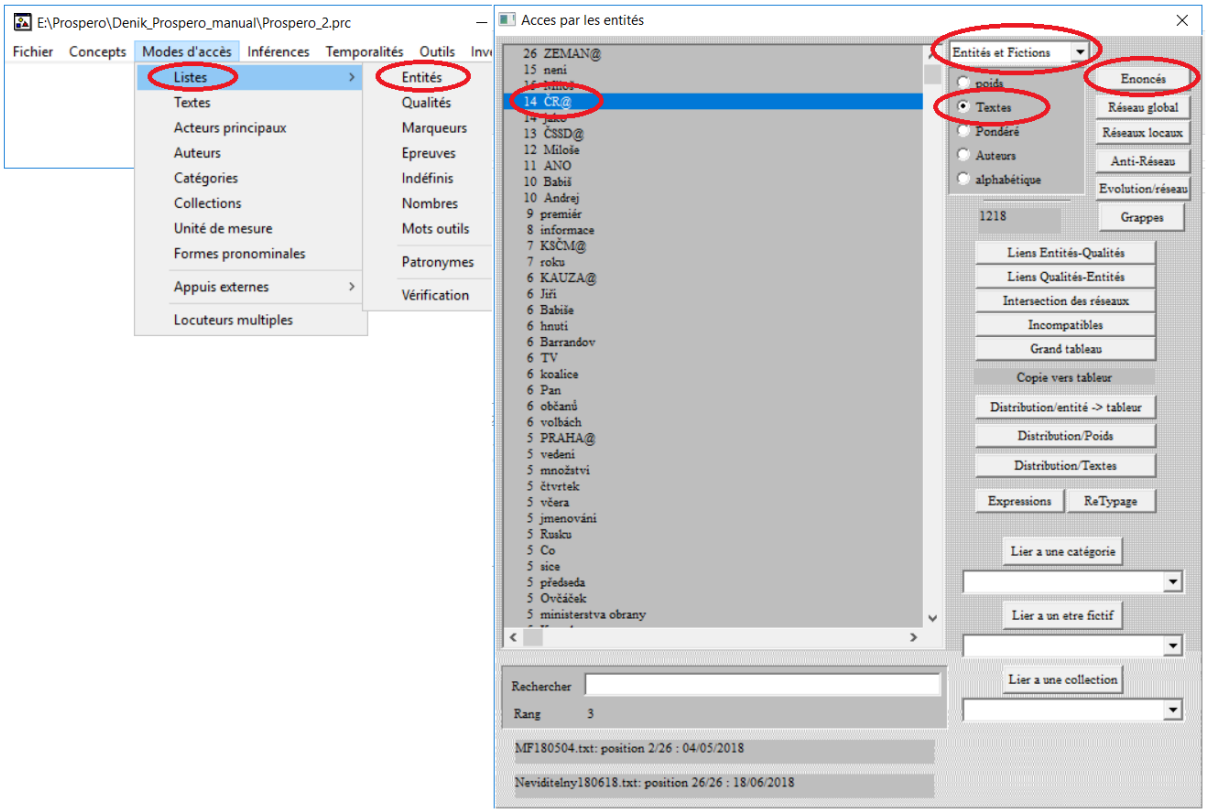
Rozdělení korpusu na dílčí skupiny textů lze provést na několika úrovních uživatelského rozhraní (UI). Kdykoliv se v UI Prospéra objeví přístup na jednotlivé výskyty hledaného jevu/výrazu/konceptu (*Acces [aux] énoncés* / *Enoncés*/ *Liste des Textes*), je zde možnost následně vytvářet dílčí korpusy.

Příklad: Vytvoření podkorpusů z hlediska výskytu konceptu ČR@

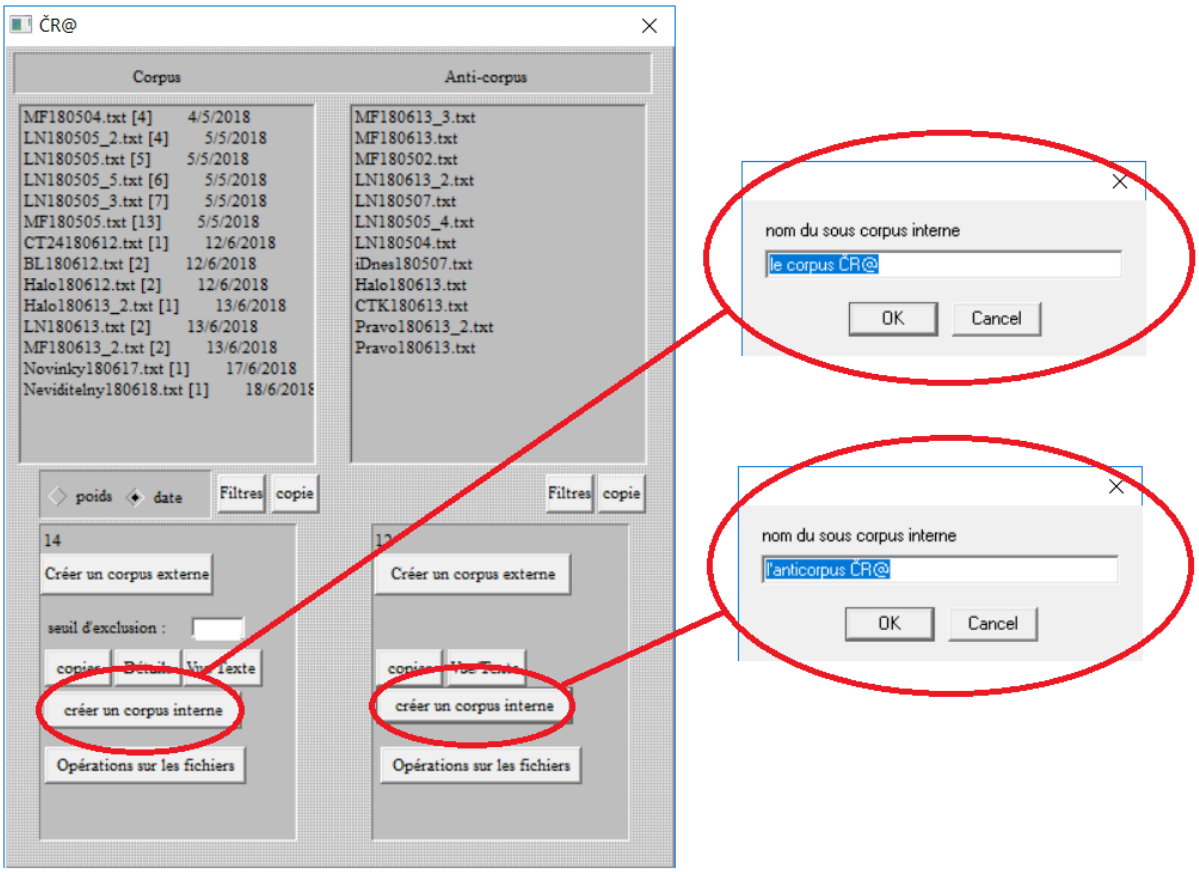
K pochopení postupu si vystačíme s jednoduchým příkladem z našeho úvodního korpusu (na obrázku výše, viz také projekt „zeman.prc"). Řekněme, že se v rámci našeho korpusu snažíme zjistit charakter výpovědí, ve kterých figuruje Česká republika jako aktér. Za tímto účelem přes funkci uživatelského rozhraní rozdělíme korpus na dva dílčí celky, tj. oddělí se všechny texty, ve kterých je přítomna velká postava (*être fictif*) ČR@, a antikorpus, který obsahuje všechny ostatní texty. Předpokladem je, že máme fiktivní postavu ČR@ již vytvořenou a zastoupenou v našem korpusu (pod záložkami *Concepts - Etres Fictifs* doporučujeme zkontrolovat dimenze (výrazy) fiktivní postavy, případně doplnit/odstranit tyto zastupující výrazy přímou úpravou souboru jazykového slovníku *cz_fictions.fic*). V rámci uživatelského rozhraní pak **vytvoření podkorpusů** odpovídá následující postup:

1. přes záložky *Mode d'accès – Listes – Entités* vybereme na seznamu „ČR@",
2. označíme volbu třídění dle typu výskytu výrazů, např. *Entités et Fictions*,
3. dále vybereme způsob třídění dle váhy výskytu v korpusu, např. *Textes*, a

4. nakonec se přesuneme přes záložku *Enoncés* na nové dialogové okno, které známe z obrázku výše.



Důležitým krokem, který následuje po vlastním rozdělení, je **uložení podkorpusů** (nejprve probereme uložení tzv. „interních“ korpusů). K tomu je třeba provést dva souvislé/zrcadlové kroky (také na obrázku níže), a sice kliknout na tlačítka *Créer un corpus interne* pod sloupcem *Corpus* (vlevo) a pod sloupcem *Anti-corpus* (vpravo). Následně se nám otevře dialogové okno k vytvoření zástupného názvu obou korpusů. Na stavový řádek doplníme libovolný název pro každý korpus a stiskneme tlačítko OK.



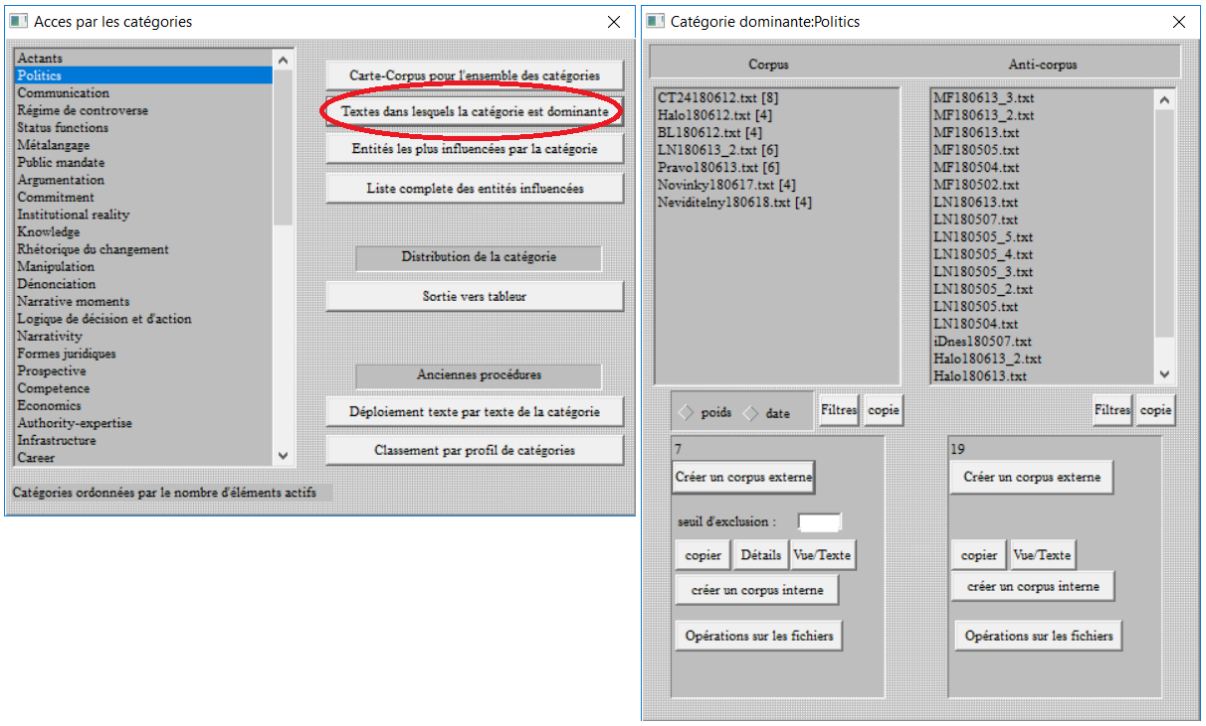
Interní korpusy nikde neexistují fyzicky, jsou pouze ve vnitřní paměti počítače. V rámci spuštěného programu drží Prospéro v paměti veškeré označené "interní" podkorpusy, tzn. naše předchozí uložené volby nepřepisuje. Toto nastavení je výhodné, potřebujeme-li získat rychlé výstupy. Dále je třeba zdůraznit, že při práci s těmito korpusy by měl výzkumník průběžně aktualizovat případné změny, tzn. v základním rozhraní projektu potvrdit změny přes *lecture des structures du corpus* a *lancer le traitement automatique*. Tvorbu tzv. externího korpusu popisujeme na konci této kapitoly.

Základní možnosti rozdělení na podkorpusy

Vytváření podkorpusů může sledovat výskyt konkrétního slova (*entité*) nebo výrazu (*Expression*), využívá se pak zejména pro sledování struktury nebo porovnávání výskytu/absence velkých postav (EF), kategorií (CAT) a kolekcí (COL). Zde se nám nabízí několik základních postupů přes záložky:

- *Mode d'accès – Listes – Entités* (obsahuje také *Etre Fictifs – EF*)*,
- *Mode d'accès – Collections – Acces Enoncés*,
- *Mode d'accès – Cat – Textes dans lesquels la Catégorie est Dominante*.

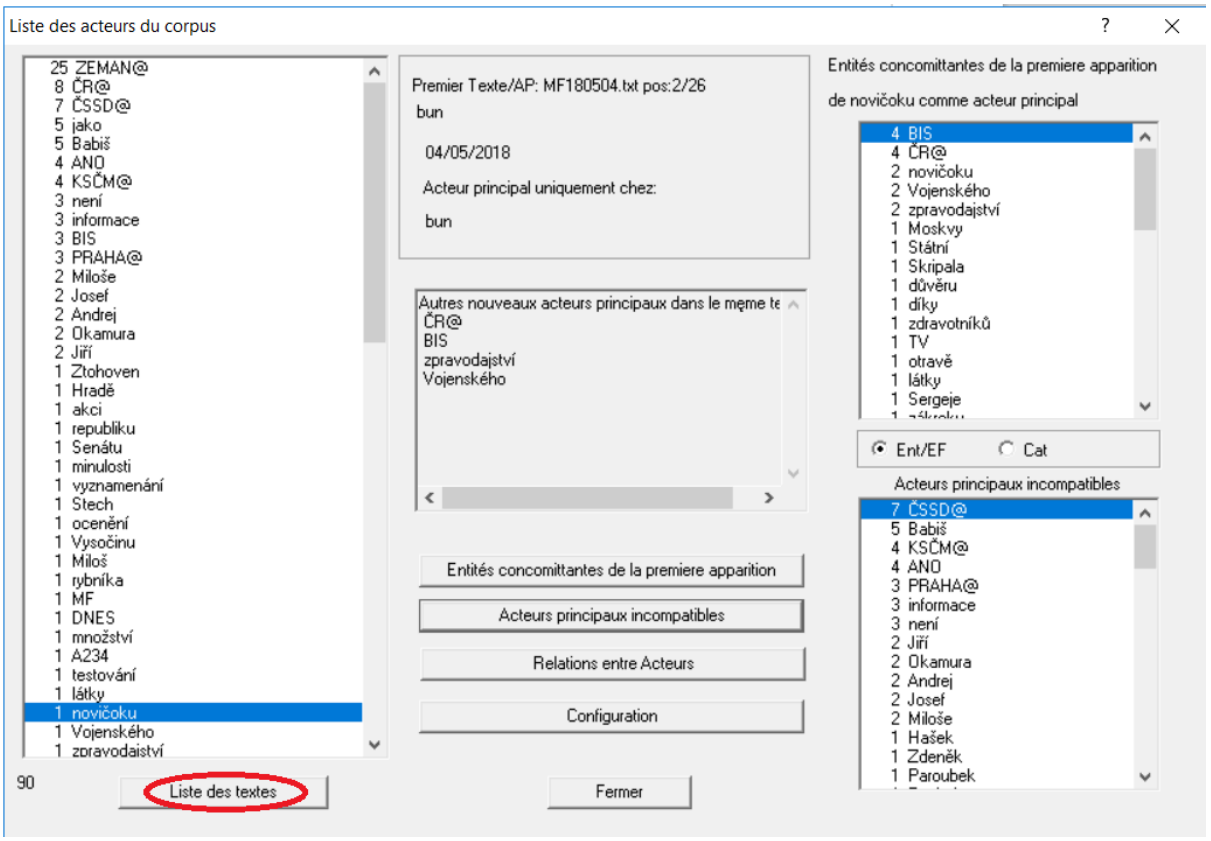
Následující obrázek ukazuje korpus a antikorpus z hlediska kategorie ***politika***.



Při volbě velké postavy (EF) nelze měřit frekvenci výskytu této postavy přímo – řešení se nabízí přes *Inférences – Configurations discursives* (nebo *Jeu d'acteurs*, viz předchozí podkapitola manuálu).

Další možnosti tvorby podkorpusů

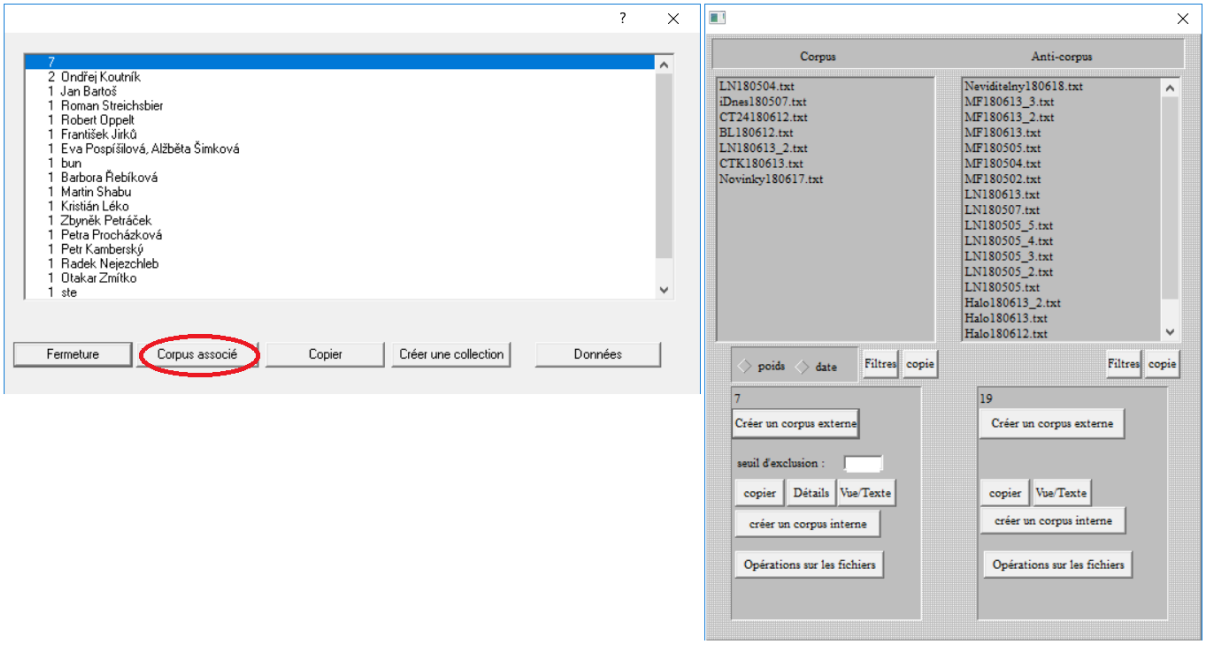
Jednou z dalších možností je rozdělit korpus na základě přehledu **hlavních aktérů** (na obrázku níže sloupec vlevo zobrazuje zastoupení jednotlivých slov a velkých postav včetně četností). Sledovaná operace se provádí pomocí příkazu *Mode d'accès – Acteurs Principaux – Listes des textes*, pro každou položku na seznamu aktérů.



Například slovní tvar „novičoku“ zde patří mezi hlavní aktéry (*acteurs principaux*) a může být sledován jako entita. Jako velká postava (*etre fictif*) by mohl být sledován v případě, pokud bychom sjednotili všechny slovní tvary (novičok, novičoku, novičokem apod).

V tomto rozhraní navíc získáme poměrně přesnou představu o subjektech, které hrají v nově vytvořeném podkorpusu důležitou roli (sloupec na obrázku vpravo nahoře obsahuje hlavní aktéry v chronologicky prvním textu o ***novičoku***): BIS a ČR@.

K dalším možnostem patří využití záznamu metadat jako rozdělovacího klíče. Přes dobře známé okno *Mode d'accès – Appuis externe* se dostaneme k seznamu metadat, např. autorům textů. V příslušné nabídce vybereme konkrétní metaúdaj (*Auteurs*) a potom stiskneme tlačítko *Corpus associé*.

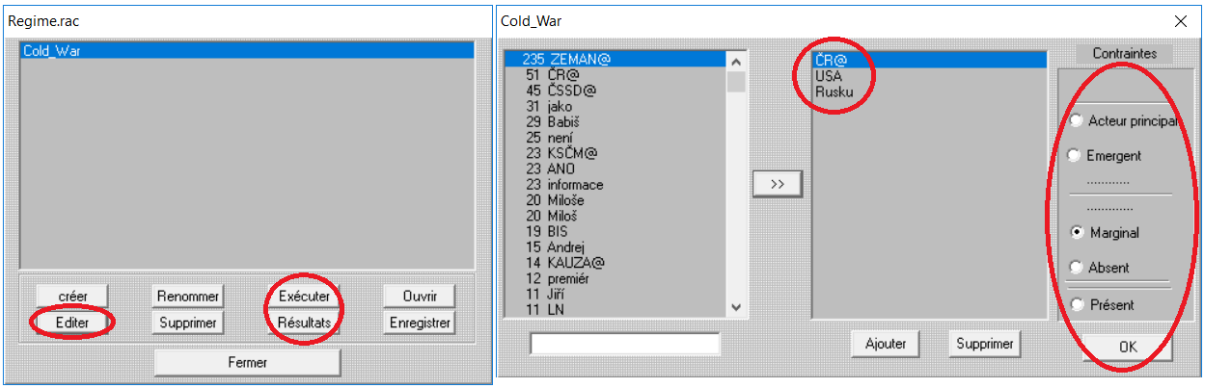


Předchozí obrázek poukazuje na jedno z možných využití rozdělení korpusu dle metaúdaje „jméno autora“, včetně možnosti oddělení textů s autory od těch, které žádného autora nemají (na obrázku výše jsme vybrali prázdný řádek).

Mnohoúrovňové vytváření podkorpusů

Chceme-li porovnávat úrovně výskytu slov, velké postavy, nebo dokonce samotné podkorpusy kombinovat mezi sebou, nabízí se následující možnosti:

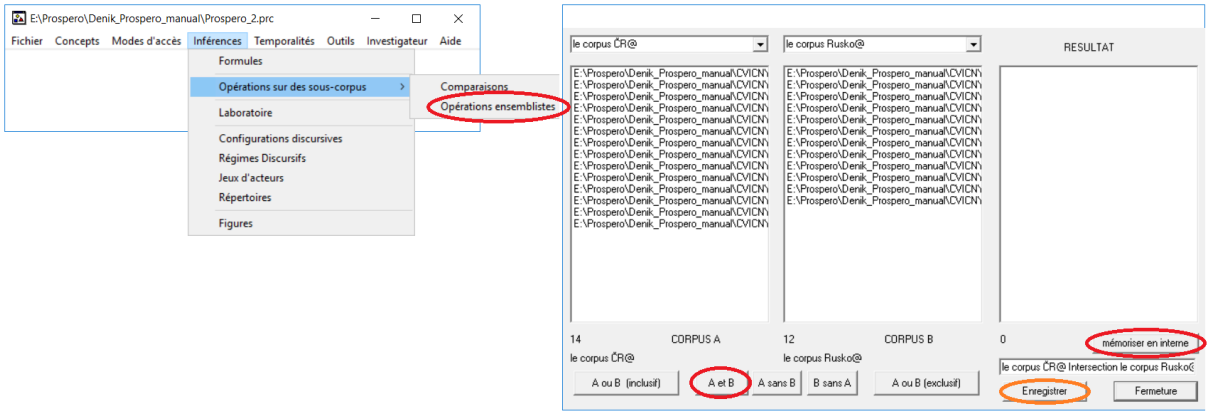
1) *Inférences – Jeux d'acteurs* sleduje souhru *entité/EF* na základě několika typů výskytu. K tomuto účelu výzkumník zavede pro sledování souhry některý zástupný název *(Créer),* a edituje typ vztahu (také na obrázku níže); Prospéro následně provede porovnání zvolené kombinace *(Exécuter)* a zobrazí výsledky *(Résultats).* Jedná-li se o často analyzovanou souvztažnost, máme možnost celou sestavu uložit a následně otevřít *(Enregistrer, Ouvrir)* při dalším spuštění programu.



Pro každý výskyt (*entité*/*EF*) zde existuje několik stupňů omezení (v pravé části obrázku výše): dominantní, emergentní, marginální, tj. tři úrovně výskytu. Přítomné (*Présent*) zahrnuje všechny tyto možnosti, zatímco absentující (*Absent*) je vylučuje. Například by nás mohla zajímat množina textů, která se vztahuje k několika výrazům (ale i *EF*): ***ČR@***, ***USA*** a ***Rusko,*** kde je výskyt jednoho hledaného výrazu okrajový (*Marginal*) a zároveň výskyty jiných výrazů (jednoho a více) jsou dominantní (*Acteur principal*), emergentní (*Emergent*), pouze přítomné (*Présent*) nebo dokonce chybí (*Absent*).

2) Další možností práce s podkorpusy je nabídka *Inférences – Opérations sur des sous-corpus – Opérations ensemblistes*. Jde o vytváření podkorpusů na základě předchozích dílčích podkorpusů „A“ a „B“. Jedná se tedy o množinové operace hledající průniky (A a/nebo/bez B) a doplňky (A a/nebo/bez B) textů mezi jednotlivými (virtuálními) podkorpusy.

Například v rámci našeho korpusu textů k danému tématu nás mohou zajímat texty, kde se píše o ČR a zároveň o Rusku nebo USA (jako jakýsi indikátor obsahu těchto vztahů). K této sestavě se dostaneme právě pomocí tzv. množinové operace (*Opérations ensemblistes*) sjednocením textů s výrazem (nebo *EF*) ČR@, Rusko a USA (také na obrázku níže). Je třeba poznamenat, že tomuto kroku předchází vytvoření (a uložení) jednotlivých podkorpusů do vnitřní paměti. Následující obrázek ukazuje texty s prezidentem Zemanem (resp. texty projektu zeman.prc) a slučuje množiny textů, kde se vyskytuje @ČR a @Rusko („*A et B*“).

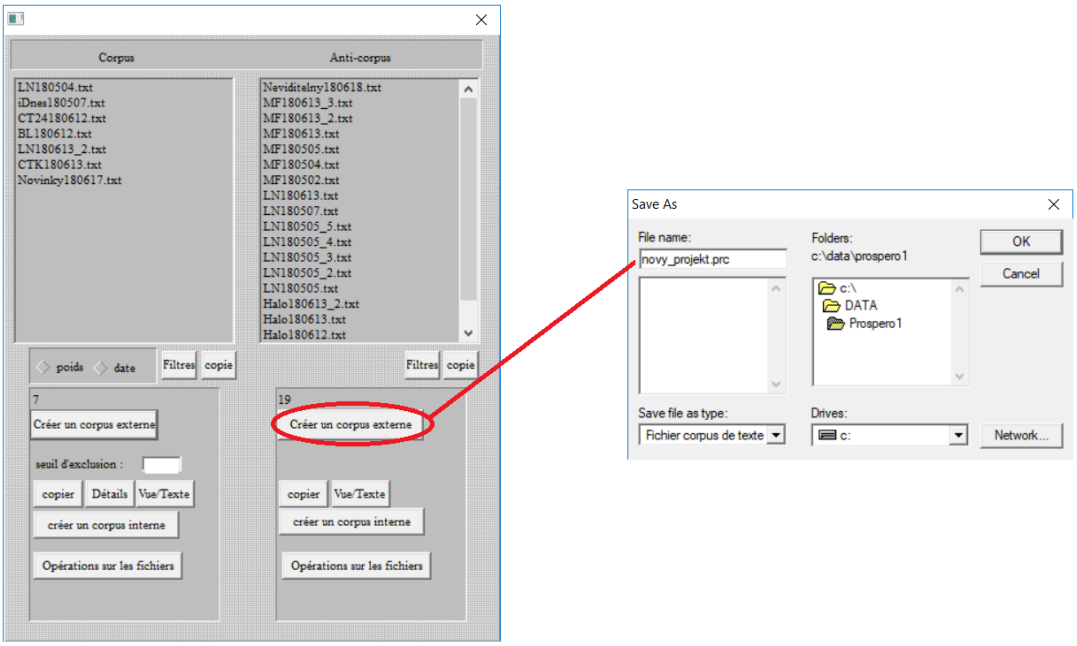


Nabízí se nám zde možnost uložení nového interního korpusu (*mémoriser en interne*) anebo uložení tohoto vztahu do nového projektu .PRC (*Enregistrer*). Tím se dostáváme k další funkci Prospéra, kterou je uložení předchozích výstupů práce s podkorpusy do nového projektu, tzv. externího korpusu.

Externí korpusy

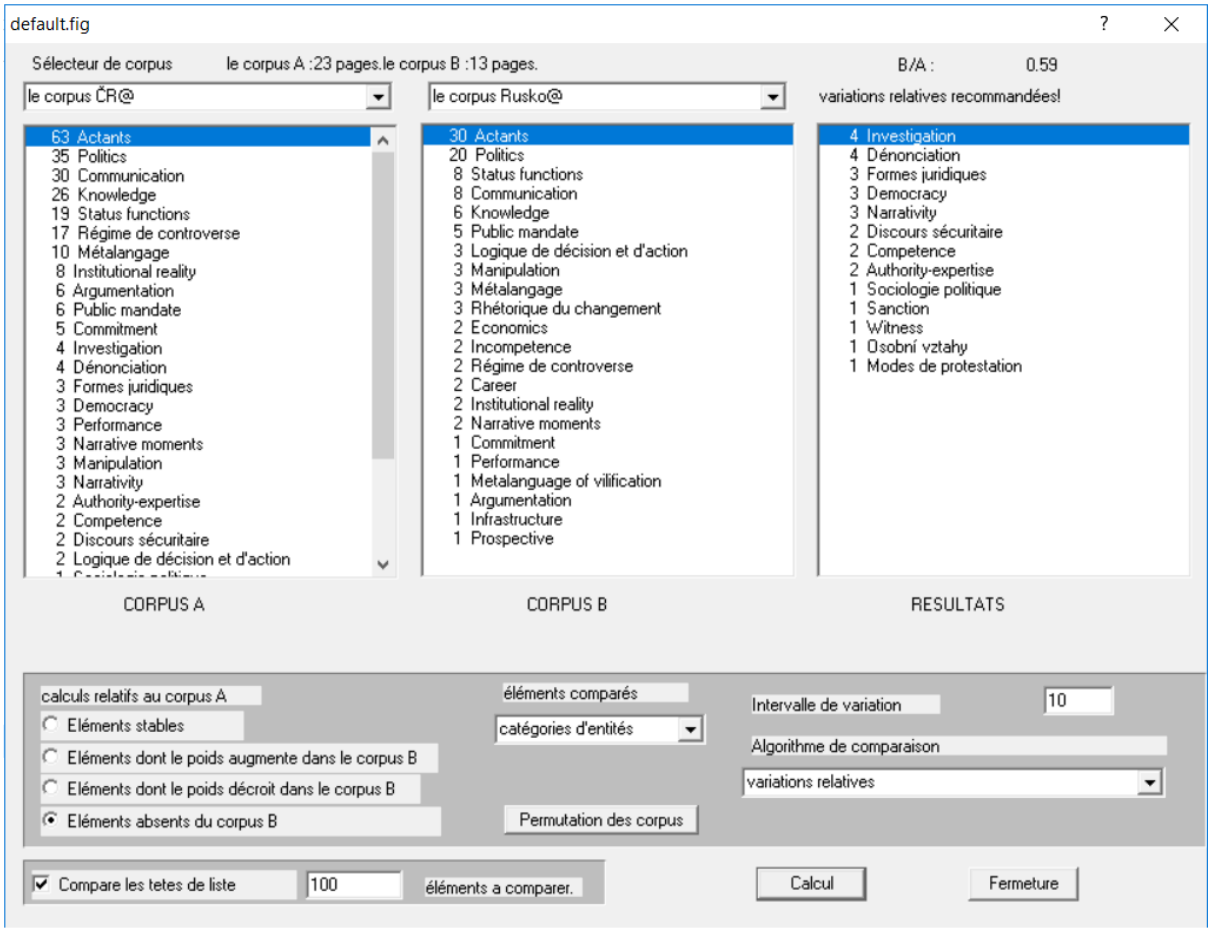
Možnosti uložit tzv. externí podkorpus se využívá zejména v případě, kdy se chceme opakovaně vracet k textům získaným na základě oddělení korpusu a antikorpusu. Jak již bylo uvedeno, Prospéro naše předchozí volby zachovává v interní paměti, která se maže při každém novém spuštění programu.

Uložením externího korpusu se zakládá nový projekt, ale již neexistuje předchozí sestava korpusu a antikorpusu (tzn. již zde nemáme volbu porovnávání – externí korpus existuje jako samostatný projekt). To nemusí být vždy nevýhoda. Řekněme, že chceme zachovat texty, které obsahují metaúdaj „autor“ jako společný projekt – nebo naopak texty bez autora nám přijdou něčím samy o sobě zajímavé. Pro každý tento korpus máme možnost uložení *Créer un corpus externe* (tento krok se provádí pro každý korpus zvlášť, také na obrázku níže).



Práce s podkorpusy (*Comparaison de corpus et anti-corpus*)

Uložením interních korpusů analýza teprve začíná. Slovy P. Trabala (2002: 17) se nám z těchto základních sestav otevírají možnosti jak navrhnout sérii „testů“, které funkce Prospéra přibližují k nástrojům textové „metrologie“ se silnou kapacitou pro objektivizaci.



Příklad je fiktivní (*EF* @Rusko), a tak jej není třeba zkoumat jako takový. Má poukázat na mechanismus porovnání „virtuálních korpusů“, které probíhá na rozhraní otevřeném přes záložky *Inférences – Opérations sur des sous-corpus – Comparaisons*. V této rovině funkce se nám odkrývá bližší pohled na dva korpusy z hlediska zvolených nastavení/kritérií. Mezi tato nastavení patří:

- volba dílčích korpusů,
- volba typu porovnávaných prvků (*entités, marqueurs, épreuves, qualités*, pojmy, osoby) apod. – na panelu uprostřed dole,
- volba typu vztahů – na panelu vlevo dole,
- vážení těchto vztahů (tzn. položek na seznamu) dle „kvantitativních“ kritérií – na levém spodním a pravém spodním panelu,
- indikace druhu algoritmu porovnávání z hlediska absolutního versus relativního výskytu.

Postup na konkrétním příkladu by byl následující:

1. Nejprve zvolíme dílčí korpusy A a B, které budeme porovnávat. V levé horní části rozhraní (na obrázku výše) se nachází dvě rozbalovací lišty, a kde je možné označit jakékoliv

korpusy, které byly uloženy – můžeme porovnávat korpusy s anti-korpusem, ale také dva různé anti-korpusy. Například označíme ze seznamu naše dva korpusy: ČR@ a Rusko@.

2. Nyní vybereme skupiny položek, které chceme v rámci korpusů porovnat. Můžeme si vybrat mezi srovnáním podstatných jmen (*entité*), jmen osob (*prénoms*), hlavních aktérů (*acteurs principaux*) anebo kategorií přídavných jmen (*qualité*), sloves (*épreuve*), příslovcí (*marquer*), ale i kolekcí. Pomocí rozbalovacího menu (na obrázku dole uprostřed) vybereme například *entités*; ty se následně objeví v prvních dvou sloupcích pro dva zvolené korpusy.

3. Nyní máme několik možností úpravy porovnání položek ze seznamu (část *Calculs relatifs au corpus A*), tzn. porovnáváme konfiguraci položek korpusů, které mezi sebou mají/nemají rozdíly. Tato nová, výsledná konfigurace se objeví ve třetím (pravém) sloupci rozhraní. Při nastavení můžeme označit rozdíl, který považujeme za významný, úpravou hodnoty proměnné, tzv. „intervalu variace". Použijeme-li přednastavený interval variace 10 (v procentech), vztahuje se tato hodnota k míře podobnosti výskytu prvků (*Elément stable*); nebo porovnání rozdílu výskytu prvků, jejichž zastoupení roste/klesá v korpusu B (vzhledem ke korpusu A). Poslední, ale neméně zajímavou možností je pak sledovat položky, které se v korpusu B nevyskytují (vzhledem ke korpusu A).

4. Dalším krokem je nastavení srovnávacího algoritmu. K dispozici jsou dvě možnosti: relativní, nebo absolutní porovnání. Srovnání zde může zohlednit poměr mezi velikostí dvou podskupin. V našem případě odpovídá korpus ČR@ 23 stranám, zatímco korpus Rusko@ je rozložen na 13 stranách všech textů, což odpovídá poměru 0,59 (tato informace se objevuje v pravém horním rohu okna nastavení). Absolutní porovnání je možno vzít v úvahu, pokud by poměr objemu/velikostí obou korpusů byl blízký hodnotě 1. V opačném případě (a to je většina případů podkorpusů) skóre jednoduše klesá/roste kvůli rozdílu velikosti dílčího korpusu. Algoritmus „relativních variací" proto dává možnost dopočítat očekávané počty (za předpokladu podobného objemu) pro kompenzaci těchto možných odchylek.

5. Tlačítkem výpočtu (*Calcul*) se celá operace spustí. Výsledky se zobrazují v pravém sloupci. Dvojitým kliknutím na některou položku v levém/prostředním sloupci snadno dohledáme položku do páru k ověření vzájemného poměru výskytu („skóre"). Po kliknutí pravým tlačítkem myši na některou položku výsledného seznamu můžeme vyhledat (*rechercher*) konkrétní výsledek (je-li např. seznam příliš dlouhý), kopírovat celý seznam apod.

Poznámka: V případě dlouhého seznamu položek, navíc s malými četnostmi položek na jeho konci, máme možnost tento seznam omezit na nejčetnější případy. Zaškrtnutím pole *Comparer les têtes de liste* s přednastavenou hodnotou 400 zobrazíme porovnání pro

prvních 400 nejčastějších. Volbou jiné hodnoty, např. 10, pak můžeme omezit seznam na několik málo předních položek.

Závěrem je vhodné připomenout, že zatímco existuje několik dimenzí porovnávání (*comparer*) korpusů na základě výskytů slov, konceptů a jevů, práce s Prospérem zde stále sleduje dva základní teoreticko-metodologické přístupy:

- porovnání tzv. „zevnitř“, vycházející z entit, jevů a konceptů získaných na základě práce s naším vlastním korpusem („corpus-driven“),
- porovnání na základě konceptů získaných z jiných projektů (teorií), od jiných autorů, které aplikujeme na náš korpus „zvnějšku“ („corpus-based“).

V tomto smyslu by komparace neměla být samoúčelnou operací, ale vstupem do další výzkumné práce (např. včetně přehodnocení toho, jakým způsobem jsme získali podkorpusy, ale i koncepty, které dimenze zvolené kategorie sledují apod).

v. Síťová (grafická) analýza propojení entit (práce s externím programem)

Program Prospéro nabízí možnost propojení výstupů s rozhraním některých externích programů, např. Pajek a Gephi, za účelem „síťové“ analýzy. Vytváření síťových map umožňuje vidět diskurzivní nebo textové struktury korpusu jiným způsobem (viz Chateauraynaud a Debaz 2009).[3] Pomocí prostorového zobrazení prvků a jejich spojení tak můžeme znásobit možnosti sociologické analýzy. Síťová analýza, kterou provádíme v Prospéru, také zde vychází z obecného principu, kdy je síť formována na základě prvku (výrazu, konceptu, jevu) „X“, a který je zároveň obklopen jinými prvky sítě, které přicházejí do kontaktu s X přes režimy výpovědí.

Základními jednotkami sítě jsou v Prospéru entity (entité). Mapa (síť) je tedy prostorový záznam (topologie) tvořený vším, co se v rámci výpovědí dostane do blízkého okolí entity. Prospéro systematicky pracuje s úplným obsahem aktérů, textů a konceptů, jimiž jsou si entity blízké, například podstatná jména (*entité*) spojují slovesa (épreuve), příslovce (*marqueur*) apod. Vždy se jedná o budování sítí na základě argumentativní logiky, kdy argumenty jsou zachyceny jako formy modalizace vztahů mezi entitami. Tyto vztahy lze navíc modelovat nejen pro celý korpus, ale i pro konkrétní podkorpus (uložený externě),

[3] Chateauraynaud, F., Debaz, J. (2009): *Des relations colatérales entre Prospéro a Pajek* [https://socioargu.hypotheses.org/141].

nebo dokonce v rámci některé kategorie a kolekce – také ty jsou množinou prvků, které lze pozorovat jako síť.

V Prospéru se objevují **tři hlavní využití** síťových map:

- k identifikaci entit, prvků a mediátorů (prostředníků) mezi a priori nespojenými světy (vygenerovaná propojenost univerza v rámci korpusu) v uvažovaném diskurzu,
- k dílčí analýze podkorpusu, konceptu (*EF*, kategorie, kolekce) vytvořeného v předchozí analýze, tzn. na základě předchozího uvažování o skupině entit,
- k hledání polarity nebo disociované sítě; stupně propojení; propojení přes některou entitu; odhalení centrálního bodu sítě anebo její periferie; změna měřítka („zoom in/out“) nebo oddálení pohledu. Prospéro tak umožňuje lokalizovat specifické sítě.

V těchto podmínkách například zjistíme, že “novičok” je spojen s „ČR“, „BIS“, „Skripalem“, „zpravodajstvím“ a „prezidentem“, což představuje několik způsobů, jak v souhrnech evokovat kauzu. Stejná analýza nám také umožňuje identifikovat, že specifická síť Skripala se skládá z „agenta“, „jedu“, „Anglie“, „Ruska“, ...

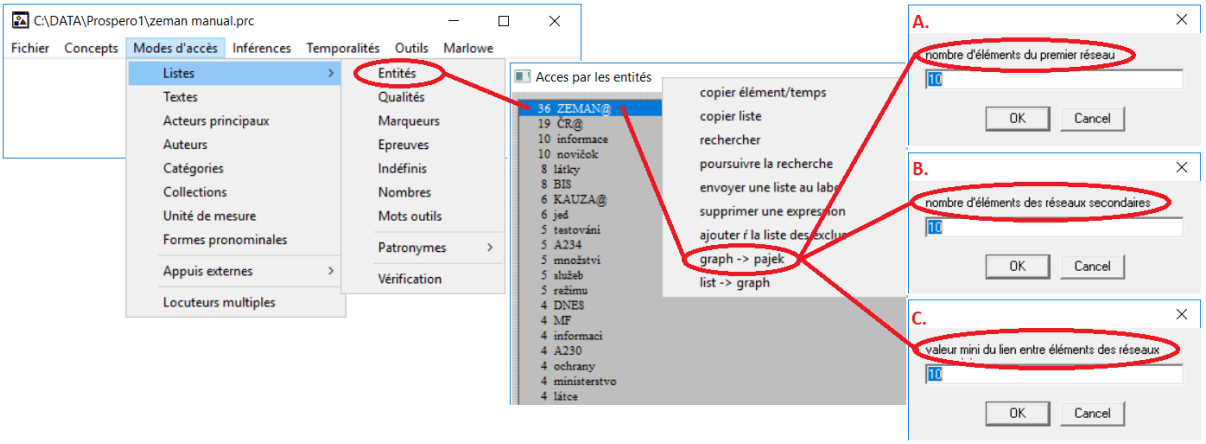
Postup vytvoření a záznam sítě: entity a koncepty

V uživatelském rozhraní přejdeme na záložku se zobrazením seznamu (*Liste – Entité*) a vybereme konkrétní **entitu/EF**. Zde na položku seznamu klikneme pravým tlačítkem myši. Zobrazená nabídka obsahuje řadu příkazů. Kliknutím na „*Graph -> Pajek*“ (také na obrázku níže) pak následuje volba třech základních parametrů sítě):

A. Nastavení počtu prvků první úrovně sítě. Prospéro omezuje výběr na spojení, která zůstávají z hlediska statisticky dominantních prvků důležitá. Tato funkce jednak umožňuje identifikovat centrální body (*sommets*), ale i prostředníky (*médiateurs*) a periferii sítě. Např. ***Novičok*** má ve zprávách korpusu množství vazeb na slovo agent, expert, jed apod., pokud jde o první desítku statisticky významných spojení. Novičok se stává (statisticky) centrem sítě a ostatní vazby představují buď jeho prostředníky, nebo periferii.

B. Počet prvků druhé úrovně sítě. Toto nastavení dále určuje velikost struktury sítě – tzn. že každý bod sítě spojený s centrem se sám o sobě stává „centrem“ druhé úrovně sítě. V našem příkladu mohou být entity „agent“ a „expert“ dále strukturované na „Anglie“, „Rusko“ apod. V jistém smyslu mohou vyjadřovat paletu vlastností každého bodu (vrcholu) sítě (nikoliv ve smyslu vlastnosti – *qualité*, a modality – épreuve, která se stává součástí algoritmu výpočtu vazeb).

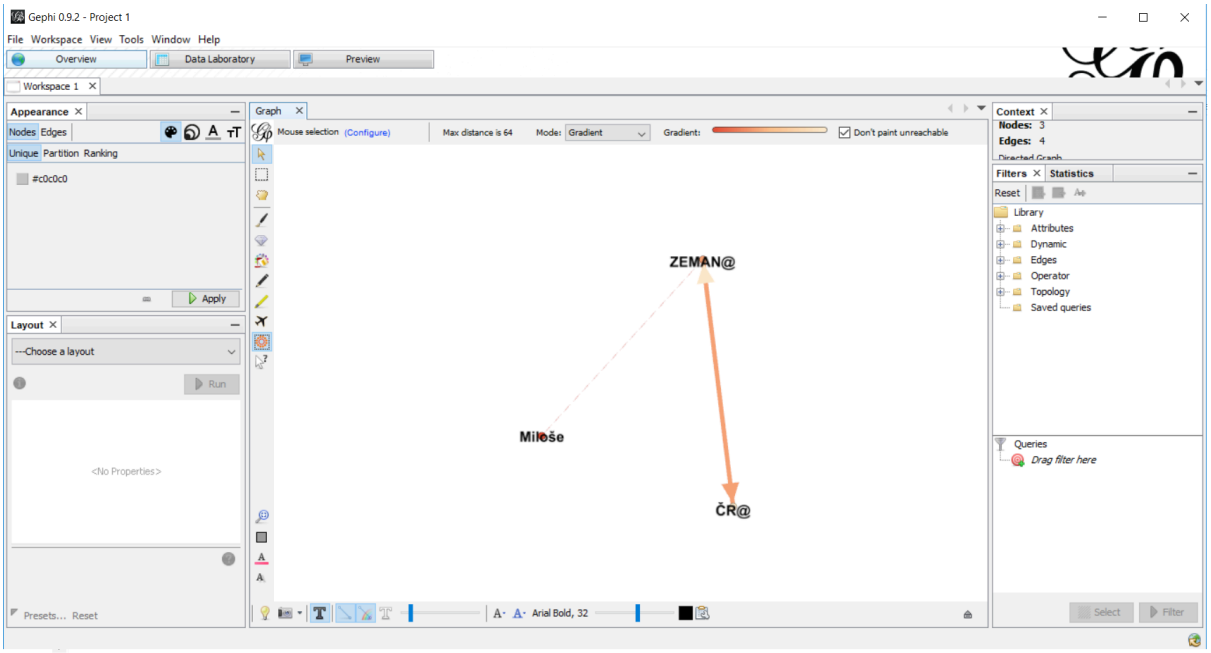
C. Minimální hodnota počtu vazeb (vektorů) mezi dvěma entitami v síti (vyšší číslo snižuje hustotu sítě/propojení, tzn. pokud nechceme zahrnout prvky v okolí entity s malou četností, nastavíme vyšší číslo).



Po volbě všech parametrů sítě se objeví okno uložení souboru ve formátu .NET. Tento získaný soubor můžeme následně otevřít v externím programu (Pajek, Gephi nebo další programy kompatibilní s formátem souborů .net).

Poznámka: Pokud nechceme měnit místo uložení sítě, musíme v instalaci Prospéra nejprve vytvořit složku „mrlw“ (tato složka, do které se sítě ukládají, nevzniká automaticky). Další možností je uložení sítě (soubor s koncovkou .net) na jiné místo.

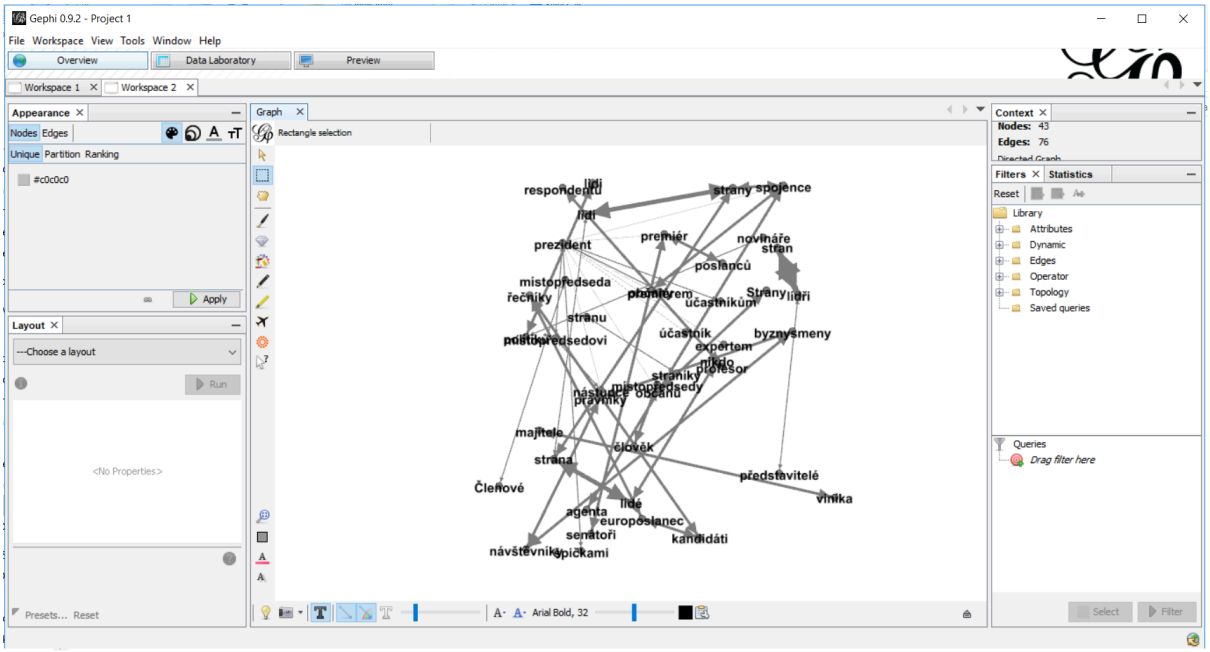
Příklad mapování entity: na základě modelového příkladu z našeho cvičného korpusu (nejprve o velikosti dvou článků) získáme jednoduchý graf (na obrázku níže vidíme zobrazení v programu Gephi). Zde je patrné, že postava Zeman@ (*etre fictif*, resp. *EF*) je propojena jednak s ***Milošem*** (výskyt *EF* je úzce propojen s entitou podstatného jména, zde konkrétně s jeho jménem) a také s ČR@ (výskyt *EF* je propojen s velkou postavou, zde konkrétně s pozicí prezidenta).



Další možností je tvorba **sítě konceptu** (velké postavy, kategorie, kolekce). Každý koncept lze převést do grafu – postupujeme přes záložky *Concepts – Catégories* – volbou konkrétní kategorie ze seznamu a pak přes tlačítko *Graphe Pajek*.

Při nastavení parametrů mapování konceptu opět postupujeme přes tři úrovně definování omezení sítě. Nejprve však musíme rozhodnout, zdali chceme omezit tvorbu sítě kolokací výpovědí v rámci konceptu (kategorie/sbírky), anebo síť orientovat také vně konceptu (tzn. potvrdíme, nebo odmítneme *calcul sur les liens internes à la catégorie*). Druhou možností je kolokace výskytu entit v rámci konceptu s jakoukoliv entitou v rámci korpusu.

Na příkladu většího korpusu si ukážeme vytváření síťového zobrazení pro konkrétní koncept, např. kategorii „aktéři“ (*Concepts – Acteurs*), a orientovaného v jeho rámci. Zde se ve výsledku jedná o jednu nedisociovanou (neoddělenou) síť.



Je na první pohled patrné, že struktura sítě a její kvalita se odvíjí nejen od struktury korpusu a volby parametrů (algoritmu), ale i od způsobu, jakým jsme koncept (kategorii „aktéři“) konstruovali. V grafu na obrázku se objevují různé entity (manuálně kódované obsahy kategorie) a některé se dokonce opakují v různých slovních tvarech (lidi, lidé, strana, stran apod.). Tato situace není chybou, ale popisuje kolokaci entit s různou váhou výskytu a jedinečností jejich vazeb v konkrétních tvarech slov („neváženým“ sloučením by síť přestala být objektivním obrazem vztahů uvnitř kategorie).

Závěrem

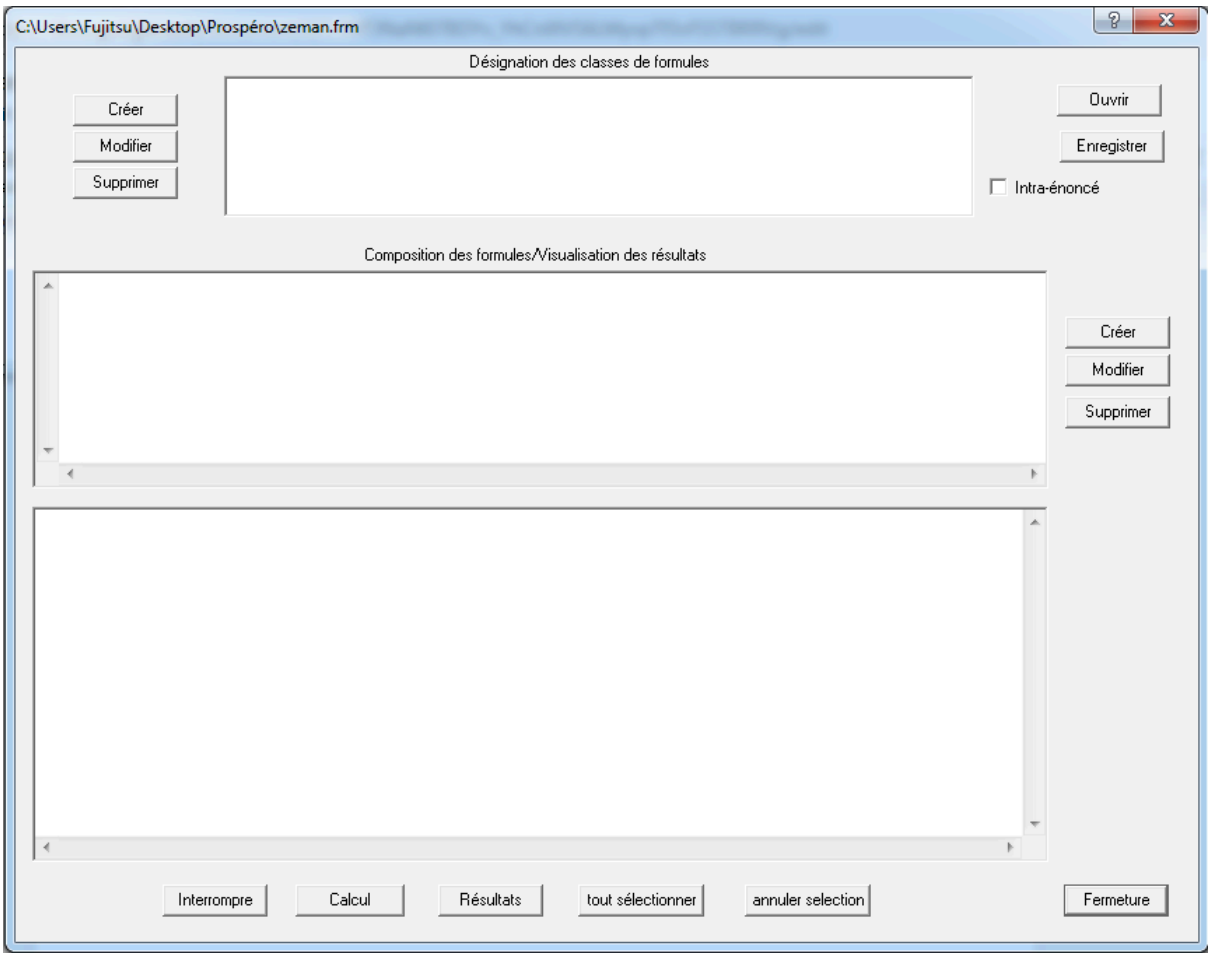
Funkce prostorového zobrazení nabízí užitečnou doplňkovou možnost objektivizace pomocí argumentačních „uzlů“ a vizualizace stupně propojení množin prvků, ke kterým tyto uzly náleží. Nicméně logika tvorby sítě nemůže plně zastoupit ani samotnou analýzu, ani dílčí operace, jako je typizace vztahů (např. tvorbu kolekcí a kategorií kódováním). V síti můžeme rozpoznat těsný vztah (např. Chateuraynaud a Debaz poukazují na prokazatelný vztah argumentace ke „změně klimatu“ s „jadernou technologií“), ale ani pak není snadné vyvodit závěry a vyšetřit vztah, aniž bychom se vrátili ke konkrétní výpovědi. To znamená, že zde nevidíme konkrétní argumenty ani jiné (lineární) struktury, jako jsou narativy (viz Chateuraynaud a Debaz 2018). Proto je třeba vždy prověřit přesnou povahu vazeb – nebo se zabývat i absencí některých entit a jejich vazeb.

vi. Formule[4]

F. Chateauraynaud popisuje funkci formulí jako „nástroj v nástroji“ nebo „program v programu“: Prospéro zde výzkumníkovi nabízí možnost překročit rámec uvažování o frekvencích, výskytech nebo sítích, které nám nabízí počítačem provedená textová analýza. Analýza se přesouvá na rovinu syntaxe nebo topologie diskurzu pomocí řady operátorů, které prozkoumávají skladbu vět, případně delších bloků textu. Vytvořené formule je doporučeno ukládat pro opětovné využití v dalších korpusech a sdílet je s kolegy pro vzájemnou inspiraci a neustálé vylepšování. Postupně tak získáme sadu formulí sloužících k úvodnímu i průběžnému zkoumání každého nového korpusu, čímž ovšem jejich přínosnost zdaleka nevyčerpáme – práce s formulemi je mimořádně tvořivá a vyžaduje důkladnou znalost korpusu i pojmových kategorií, protože jde v první řadě o umění klást dobré výzkumné otázky. Tato funkce pracuje se zobecňováním, klíčovými výrazy, logickými operátory i schopností správně sestavit zadání formule jak sémanticky, tak formálně.

Při práci s formulemi využijeme slovní druhy i kategorie. Vše ukážeme na konkrétních jednoduchých příkladech z našich dvou korpusů. Na konci této podkapitoly pak uvádíme tabulku, která shrnuje rozsah možností tvorby formulí. Rozhraní, s nímž budeme pracovat (otevírá se přes záložky: Inférences - Formules):

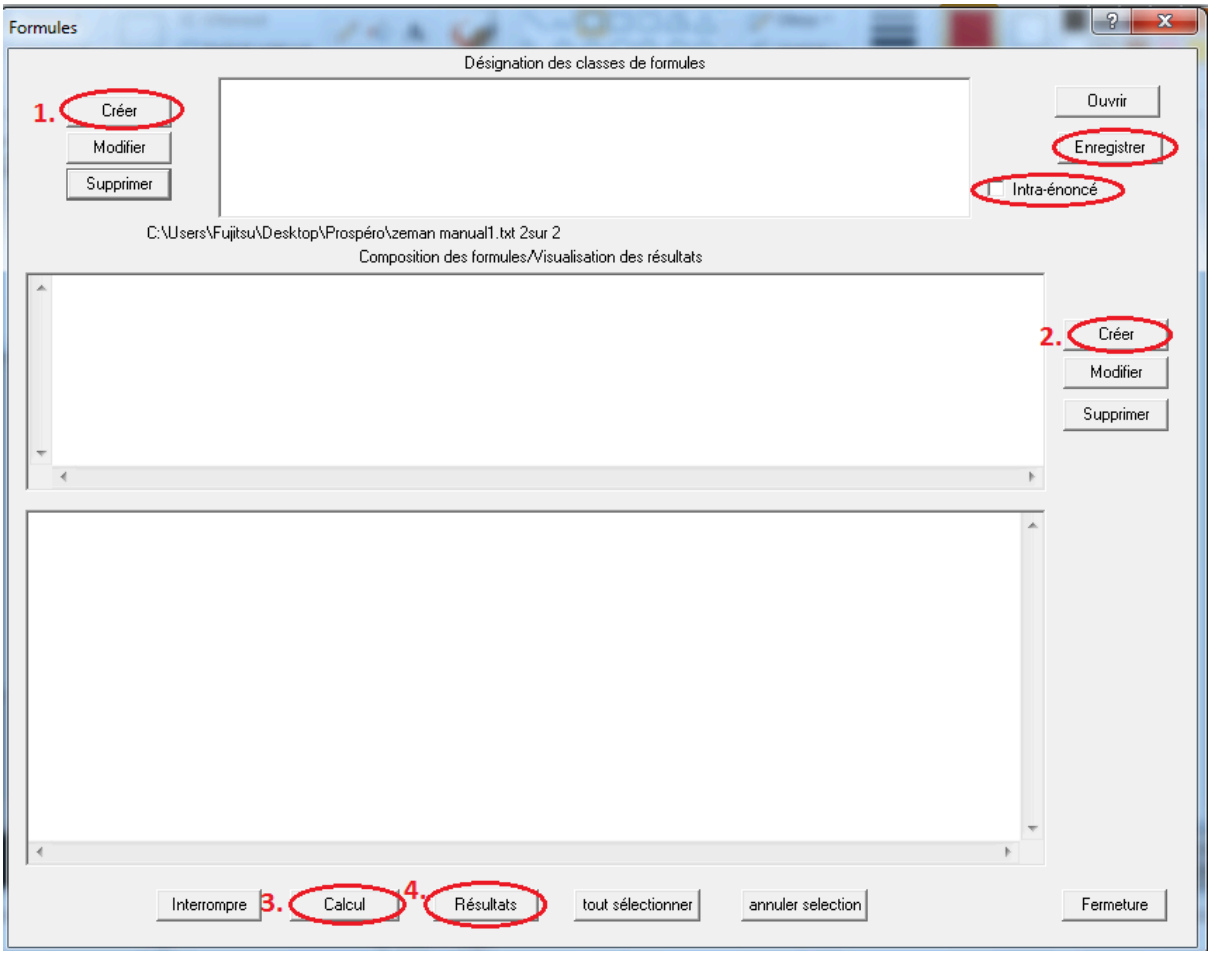
[4] Jde o vzorce v matematickém smyslu skupiny symbolů a operátorů.



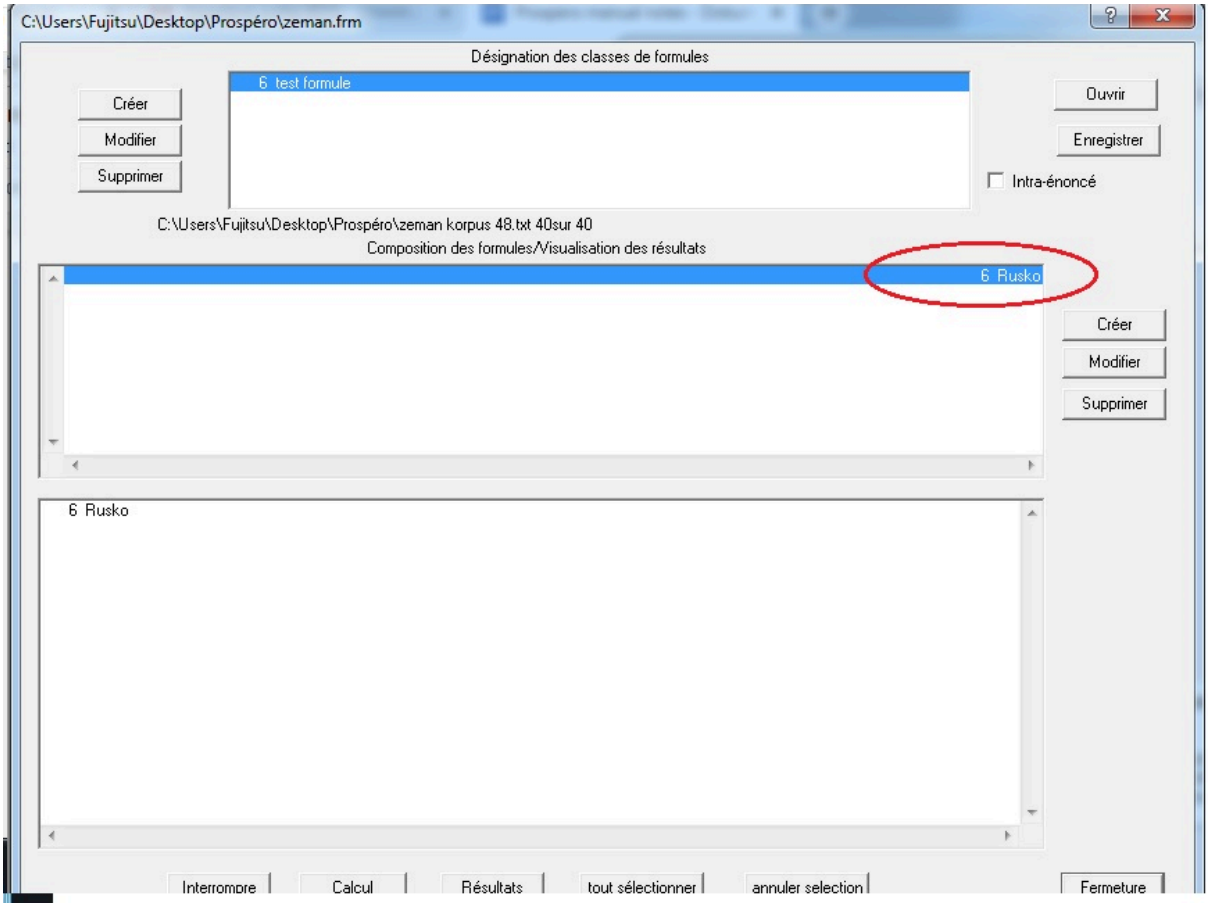
Při prvním otevření je pole prázdné (vždy je nutno nejprve provést otevření a načtení korpusu). Začněme konstrukcí nejjednodušší formule: vyhledáním jednoho slova bez jakýchkoliv dalších logických operátorů (hledání lze provést přes *Outils – Recherche par préfixe/suffixe*, jak již víme – u práce s formulí se však posouváme k analytické rovině využití vyhledaného slova).

Nejprve si s pomocí tlačítka *Créer* (vlevo nahoře) formuli založíme a pojmenujeme, například „test formule“; tlačítkem *Modifier* poté příp. upravujeme a *Supprimer* název

formule smažeme. Druhé okno slouží pro samotný obsah (resp. algoritmus) formule.



Zde pomocí tlačítka *Créer* (vpravo uprostřed) definujeme, např. zde zadáme slovo ***Rusko***. Tlačítka *Calcul* a *Résultats* poté spustí hledání nebo zobrazí jeho výsledek. Získáme tento výsledek:

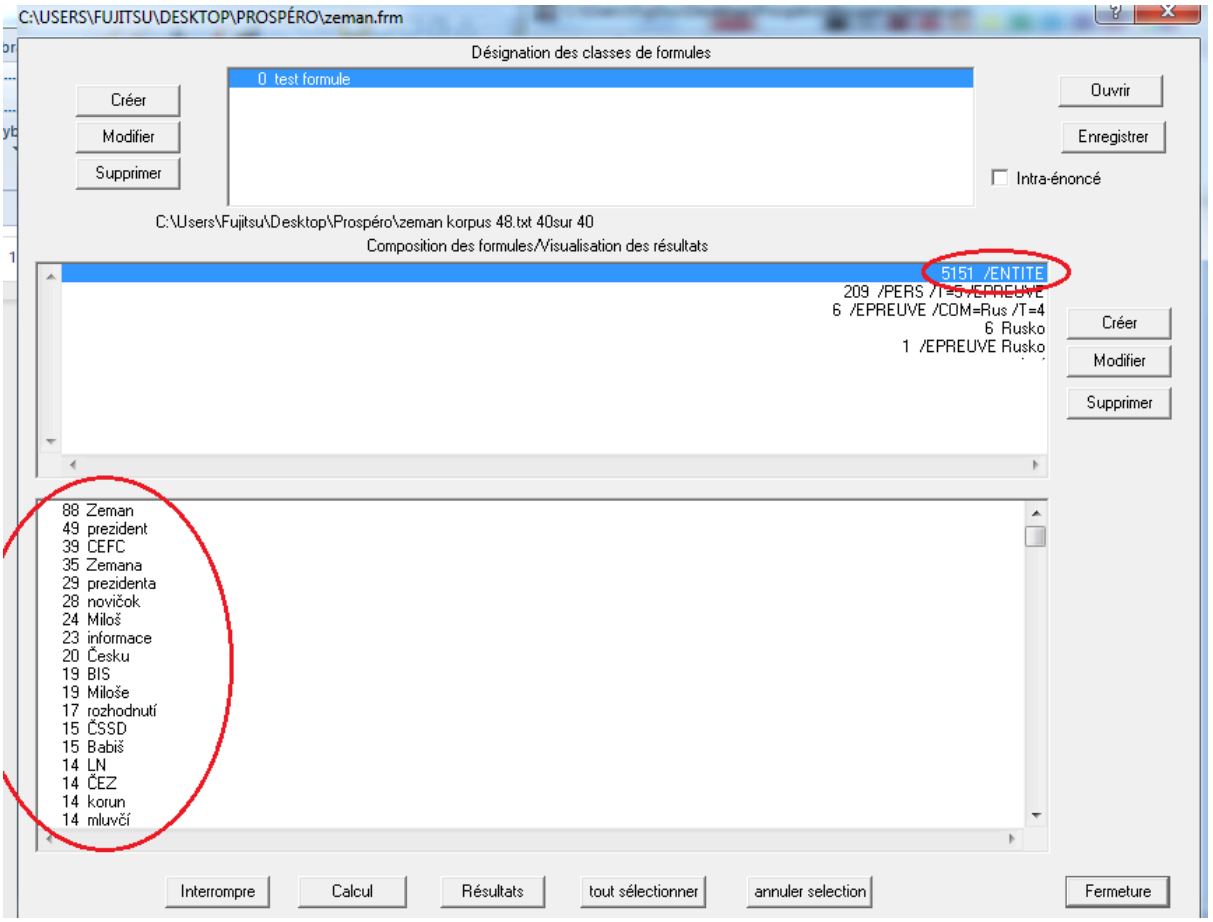


Kliknutím na řádky se zadáním formule a jejím jménem zjistíme počet výsledků – jejich celkový počet v horním okně, počet pro každou variaci formule ve středním a pro každý výsledek v dolním. Jinými slovy, v našem případě máme jen jednu variaci a jeden výsledek, který se v korpusu vyskytuje šestkrát.

Vytvořené formule ukládáme na libovolné místo (*Enregistrer*); soubory s uloženými formulemi mají koncovku .frm (typ souboru je při ukládání označen automaticky). Při každém dalším otevření Prospéra je vkládáme přes *Ouvrir*. Formule se standardně ukládají vždy k danému projektu, lze je ale kopírovat i do složek jiných projektů, případně si vytvořit společné místo pro všechny naše formule. Pozor, nově uložené nebo upravené formule se přemazávají přes formule již jednou uložené. A pokud potřebujeme kupř. ověřit správnou podobu názvu kategorie, okno formulí zavřeme; rozpracované formule nezmizí – zmizí jen, když zavřeme projekt (*Effacer projet*) nebo Prospéro.

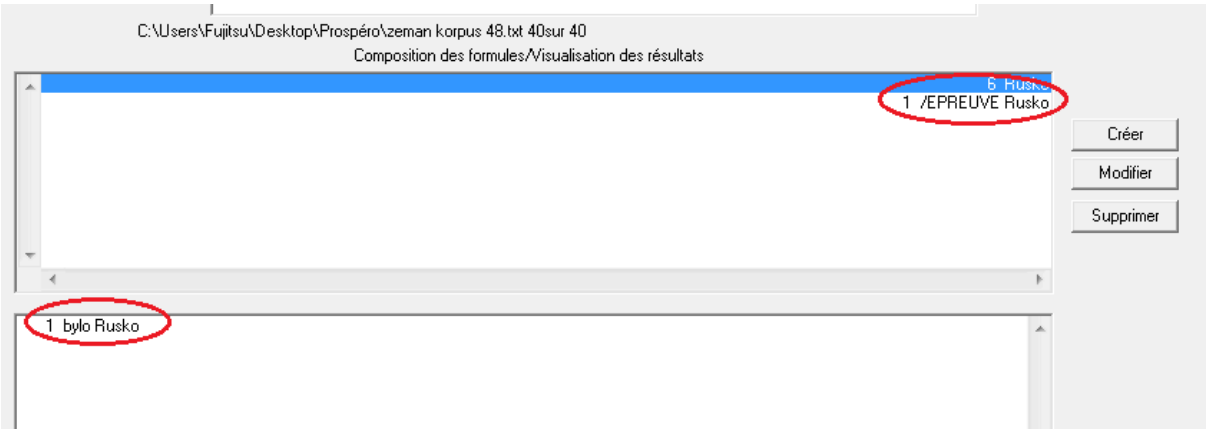
Formule s logickými operátory

Silná stránka práce s formulemi, která ji posouvá nad běžné vyhledávání slov, je možnost pracovat se zkratkami (zástupnými výrazy) a logickými operátory. Například máme možnost vyhledat všechny entity v našem korpusu:

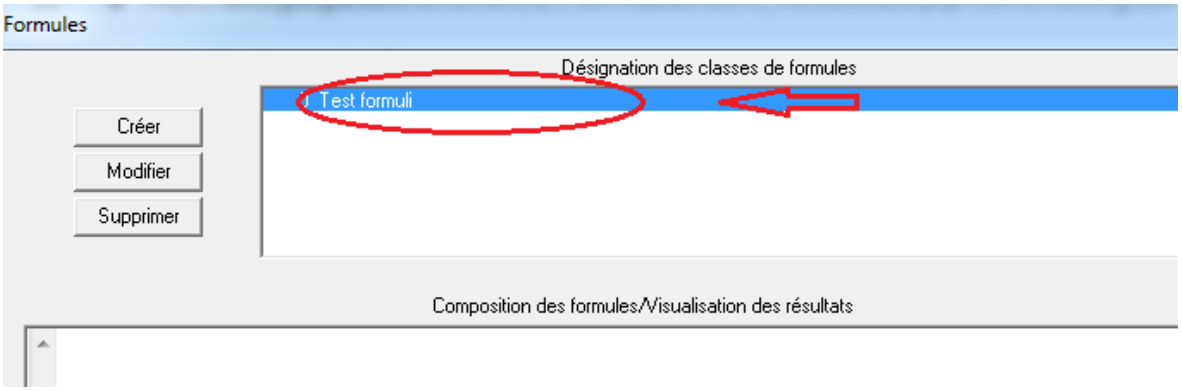


Pod stejnou skupinu formulí („test formule“) nyní uložíme další formuli, prostřednictvím níž budeme vyhledávat kombinaci jakéhokoliv slovesa (épreuve) s podstatným jménem „Rusko“. U spojení více zástupných výrazů uvnitř formule je nutné před jakýmkoliv logickým operátorem vždy psát znak **/** (lomítko). Celá formule je pak ve tvaru:

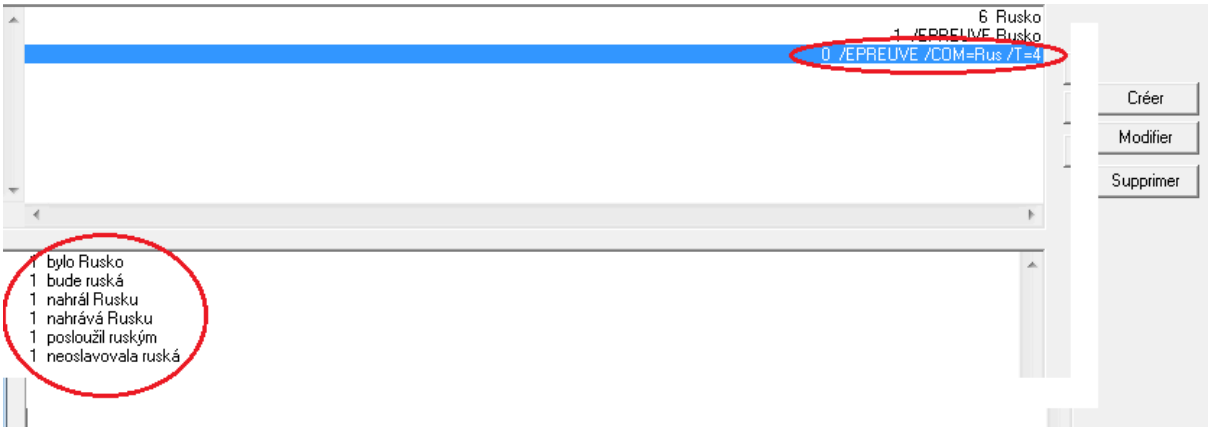
/EPREUVE Rusko



Abychom získali podobu výsledku uvedenou na předchozím obrázku, je při úpravě (*Modifier*) či vytvoření (*Créer*) formule třeba kliknout o úroveň výš, aby došlo k aktualizaci a přepočítání výskytů vyhledávání:

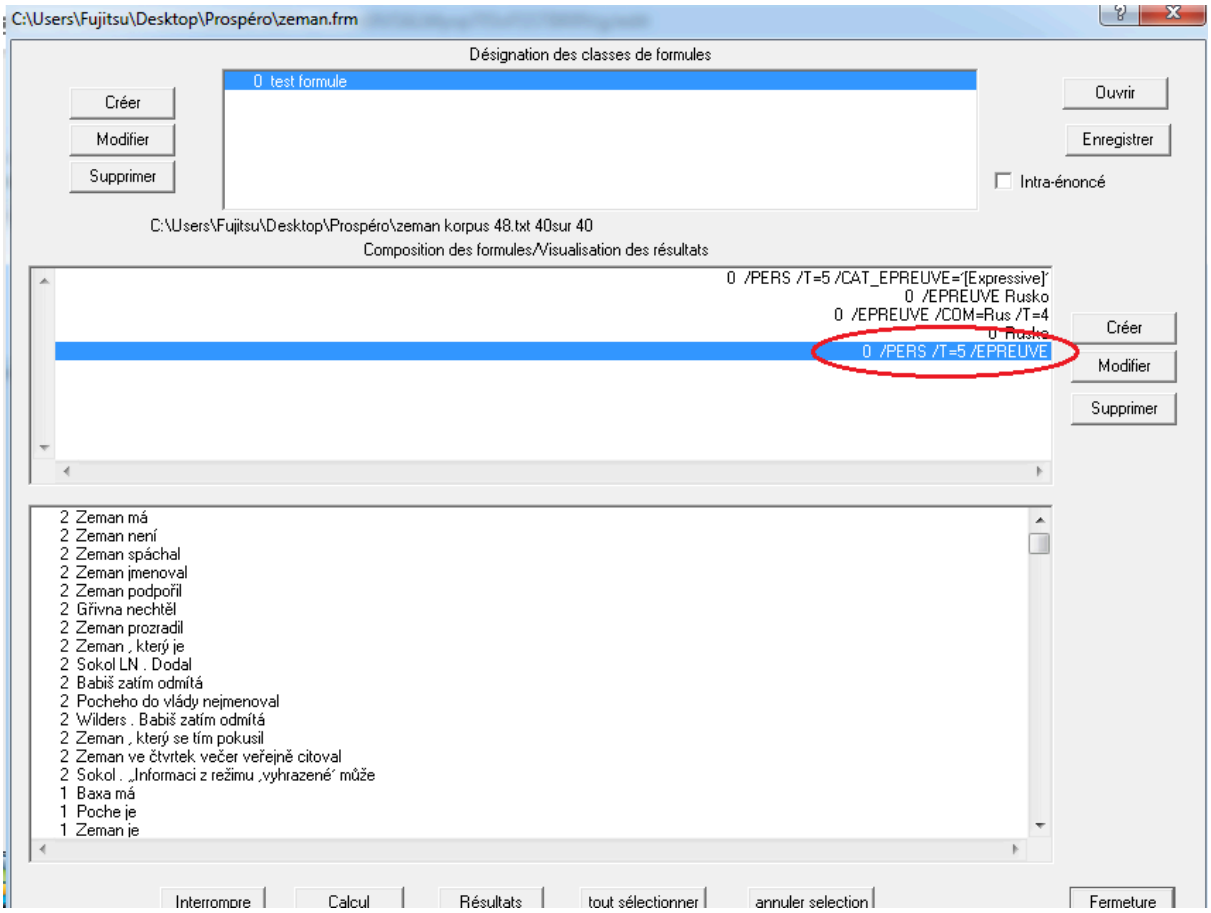


Nyní buď upravíme (*Modifier*) stávající formuli, či vytvoříme formuli novou, např. úpravou zjistíme jakékoliv sloveso, které se vyskytuje spolu se všemi slovy začínajícími na ***Rus-*** (nemusí jít tedy již jen o Rusko). Zároveň uvidíme ještě další (až 4) slova, která následují:

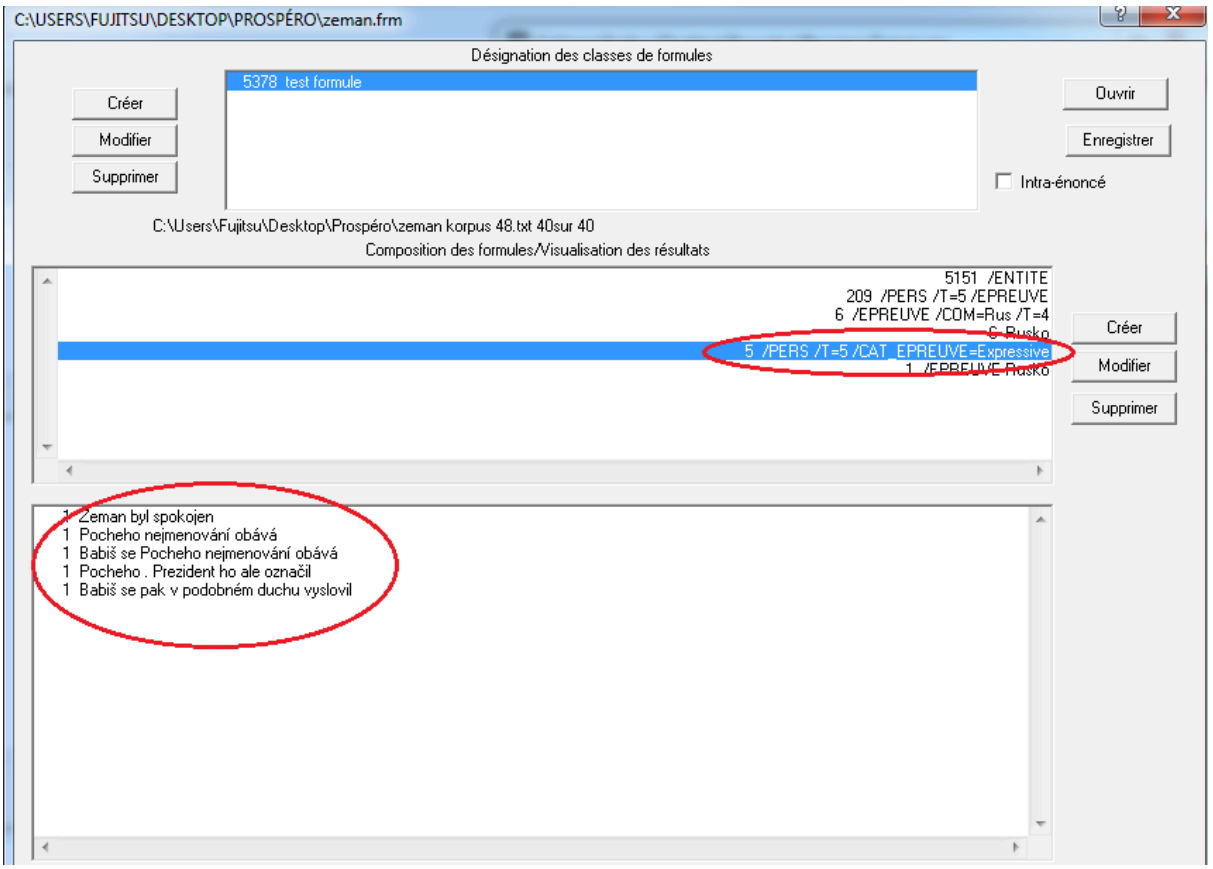


Zde se začíná objevovat smysl celé funkce ale i složitost práce s formulemi. Prospéro balancuje nad jazykovou a sociologickou analýzou, kdy jsme pomocí formule získali spojení ukazující "činný" a "trpný" režim. Entité je původcem děje: *X nahrál Rusku*. Entité je cílem děje: *Y bude ruská*. Zacílit jen na jeden z nich (pokud bychom chtěli) je prakticky nemožné (k tomuto účelu bychom museli Prospéro naučit rozdělovat režimy např. přes tvorbu některé kategorie) - zároveň se nám odkrývá část diskurzivního okolí zkoumaného jevu.

Naše vyhledávání může sledovat jiné obsahy nebo způsoby výpovědí, které dále upřesníme za pomoci nové formule. Zajímají nás jména osob spojených s nějakou činností, tedy slovesem, kde mezi nimi může být až 5 jiných výrazů:



Hledání ještě dále zpřesníme. Zajímají nás slovesa (épreuve) z konkrétní kategorie, např. *expressive*:



Možnosti kombinování a řetězení příkazů pro hledání v korpusu pomocí formulí jsou téměř neomezené. Tabulka níže obsahuje jejich úplný seznam.

Repertoár pro zápis formulí

/ENTITE	Jméno
/MAJENT	Jméno začínající velkým písmenem
/EPREUVE	Sloveso
/QUALITE	Přídavné jméno
/MARQUEUR	Příslovce, modalizátor
/MO	Pomocné slovo
/NBRE	Číslovka
/PRENOM	Křestní jméno, titul
/PERS	Celé osobní jméno
/ELEMENT	Jakékoliv jméno, sloveso, přídavné jméno či příslovce

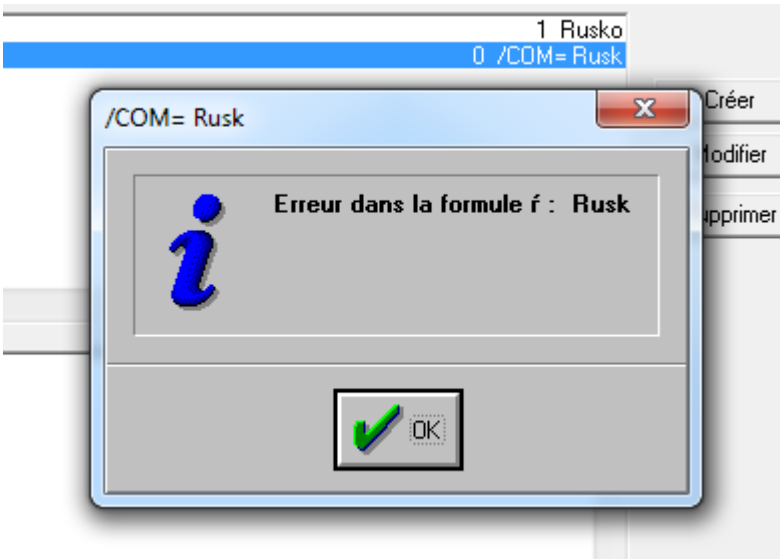
/EF	*Etre fictif*
/EF=’jméno osoby’	
/COL	Kolekce
/COL=’název kolekce’	
/CAT	Kategorie
/CAT_ENTITE=’název kategorie’ /CAT_EPREUVE=’název kategorie’ /CAT_QUALITE=’název kategorie’ /CAT_MARQUEUR=’název kategorie’	
<u>Pozor: Ve výše uvedených formulích uvádíme uvozovky jen, pokud jsou názvy víceslovné</u>	
/COM=jedno nebo více písmen	Slovo/výraz začínající na...
/TERM=jedno nebo více písmen	Slovo/výraz končící na...
/RAC=jedno nebo více písmen	Slovo/výraz obsahující uprostřed...
/CT=jedno nebo více písmen	Slovo/výraz začínající na… a končící na...
/T=číselná hodnota	Interval od 0 až do uvedené hodnoty (nemá být první ani poslední prvek formule, protože specifikuje mezeru mezi dvěma dalšími prvky)
/P=číselná hodnota	Hodnota stanovená na právě x slov
/BLA	Všechny znaky mezi dvěma jinými součástmi formule
/CF (*Classe de formules*)	Formule, která se dá umístit uvnitř další formule (hodí se např. tam, kde potřebujeme seskupit několik slov nebo kategorií pro opakované včlenění do našeho hlavní formule)

Obecné zásady vytváření a zápisu formulí

- Vyhledávaná slova nesmí být napsána velkými písmeny.
- Logické operátory naopak musí být napsány velkými písmeny a bez francouzských diakritických znamének; toto neplatí pro samotné názvy kategorií, které zadáváme do příkazu, např. Expressive (celé znění /EPREUVE=Expressive
- Před logickým operátorem vždy píšeme / (lomítko).

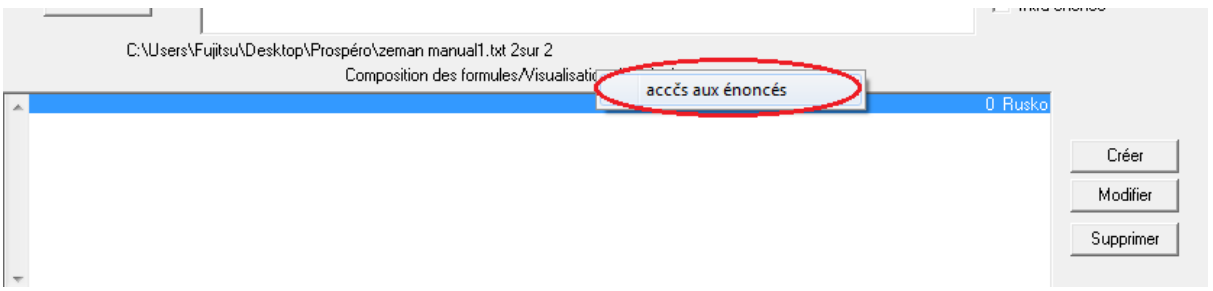
- Víceslovné výrazy je třeba dát do jednoduchých uvozovek.
- Kolonka *Intra-énoncé* – při zaškrtnutí bude vyhledáván daný jev jen v rámci jedné věty.
- Věta končí tečkou a tu čte Prospéro jako jeden znak – vykřičníky a otazníky nejsou považovány za konec věty, a proto mohou být, jakožto často významné enunciační znaky, součástí formulí i při zaškrtnutí kolonky *Intra-énoncé*.
- Při udání počtu znaků ležících mezi dvěma prvky, např. T=3, budou vyhledána spojení začínající prvním a končícím druhým prvkem s 0, 1, 2 anebo 3 slovy mezi nimi (kde „slovo“ může být punktuační znak anebo *expression*).
- Mezi jednotlivými výrazy ve formuli píšeme mezeru.

Při jakékoliv chybě se objeví hlášení (které odklikneme a formuli opravíme):

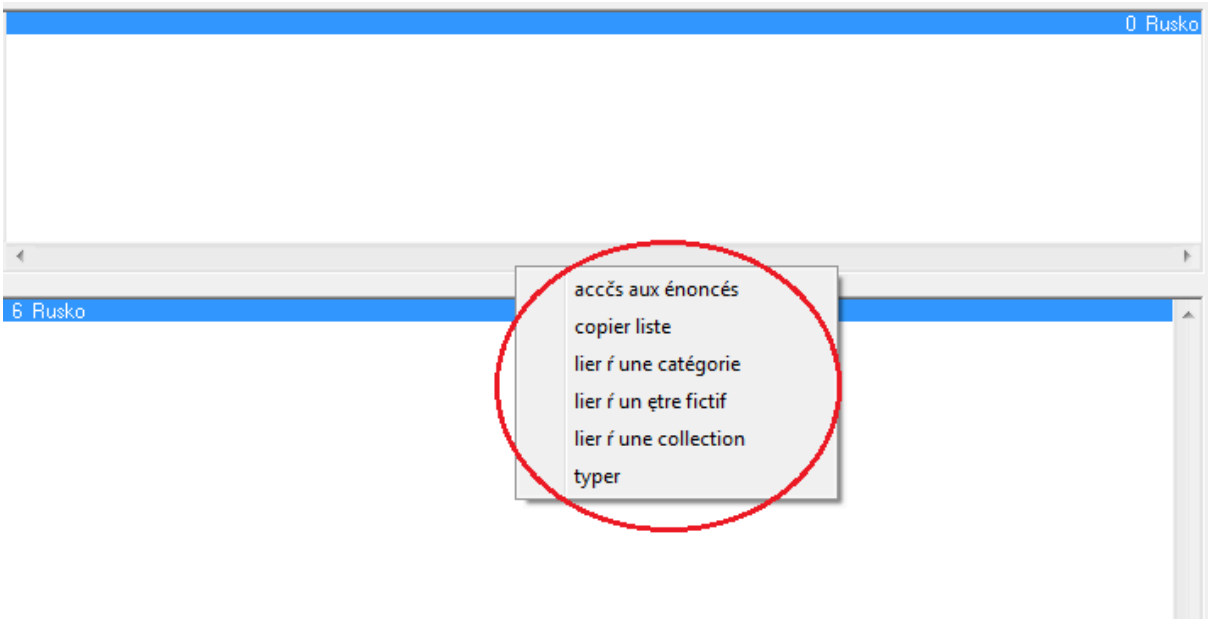


Další možnosti práce s výstupy formulí

Na rozhraní okna výstupů (dolní část rozhraní okna formulí) lze provádět některé další úpravy. Kliknutím na znění formule pravým tlačítkem lze prostřednictvím *Accès aux énoncés* přejít k jednotlivým výskytům jevu:

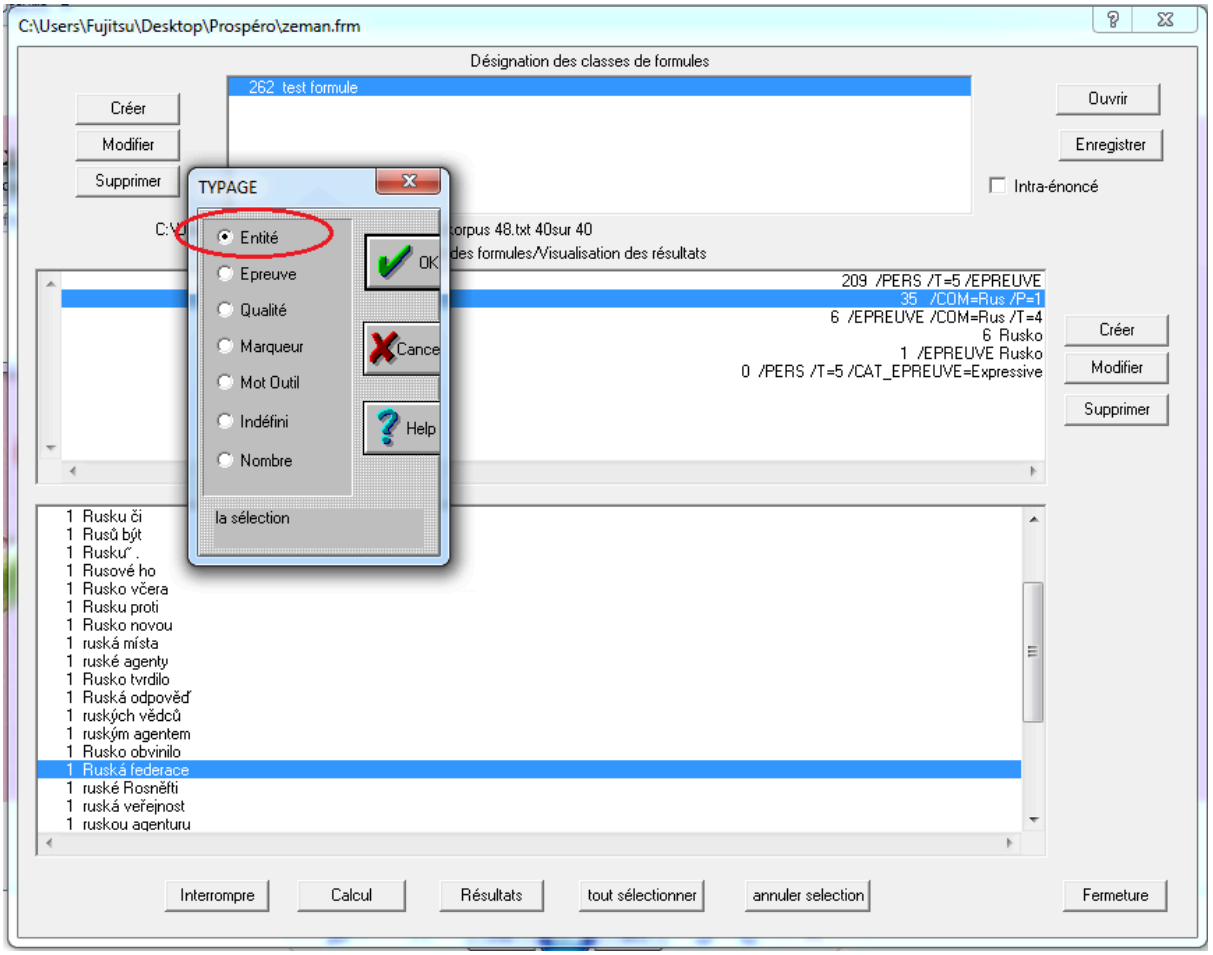


Stejně tak lze opět přes výsledky vyhledávání ukládat výstupy do kategorií, měnit klasifikaci slovního druhu apod.:



Nalezené výsledky jsou vždy řazeny dle počtu od nejvyššího. Pomocí výstupů tak lze rozpoznat častý jev a je vhodné přejít ke klasifikaci slova/výrazu, resp. pokud je to žádoucí z hlediska jazykového zařazení (typer) nebo výzkumného záměru (catégorie/être fictif/collection).

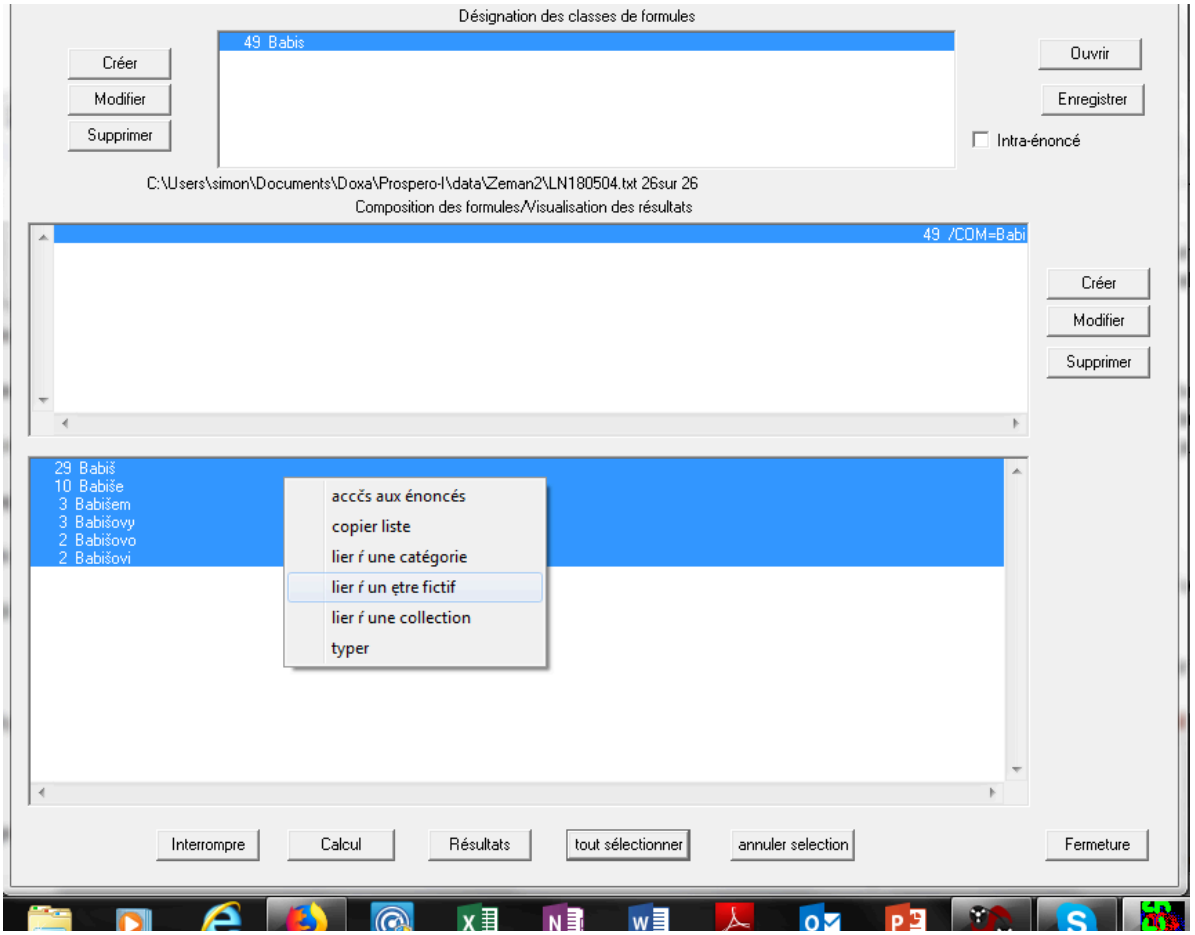
S pomocí formulí lze nalézt a vytvořit i libovolný nový frazém (*expression*).



Formule lze rovněž použít k vytvoření nového *Etre fictif*. V seznamu *entités* je čtvrté nejčastěji se vyskytující slovo ***Babiš***, a tak bychom například chtěli sjednotit všechny jeho pády a další odkazy na slova zastupující Babiše, tzn. aktéra/velkou postavu (např. ***prozatímní premiér, designovaný premiér***) do jednoho záznamu *EF*, podobně jako existuje *EF* pro ***Miloše Zemana***.

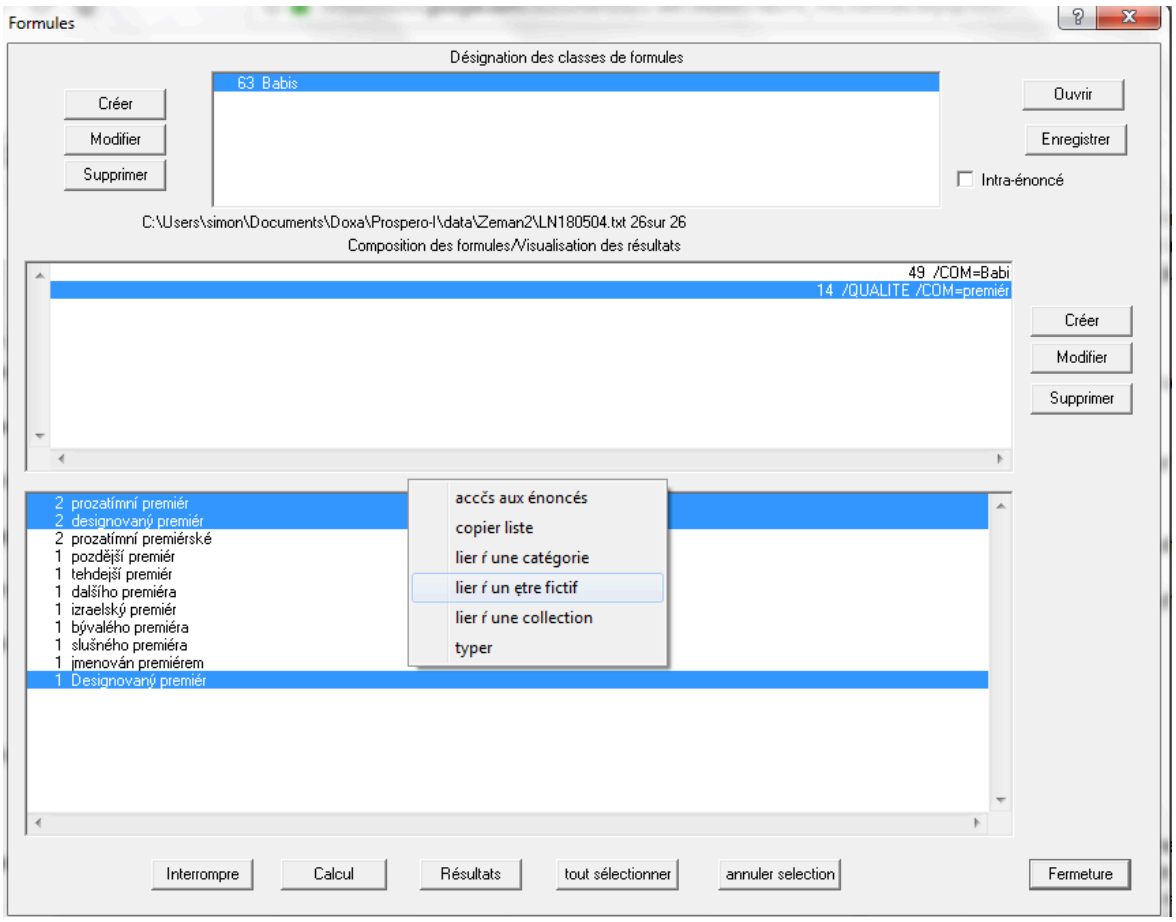
Postupujeme tak, že nejprve zachytíme všechny pády jeho příjmení formulí:

/COM=Babi



Následně vybereme všechny výsledky (*tout sélectionner*), klikneme pravým tlačítkem a zvolíme možnost *lier à un etre fictif* (dále postupujeme již známým způsobem). Na dalším řádku také vytvoříme formuli schopnou zachytit všechny kvalifikace označení „premiér“:

/QUALITE /COM=premiér



Pokud si nejsme jisti, které z výskytů označují Andreje Babiše, a které ne, pravým tlačítkem a volbou *accès aux énoncés*, můžeme zkontrolovat celou větu nebo text.

Máme-li jasno, které všechny výrazy chceme spojit v novém *etre fictif*, kliknutím na položky vybereme výrazy a opakujeme postup pro spojení s velkou postavou.

Příklady dalších formulí

Nyní ukážeme další, již vytvořenou formuli, která je součástí instalačního balíku:

Otevřeme formuli „Titulky“. Ta bude správně fungovat pouze za předpokladu, že jsme v korpusu označili titulky dvěma hvězdičkami před a po každém titulku, případně jednou hvězdičkou před a po každém mezititulku (cvičný korpus už je takto upravený - viz úpravy textů).

Formule “Titulky” umožňuje v první řadě získat seznam všech titulků (nadpisů) a mezititulků (podnadpisů) v korpusu. Nastavení rozsahu výsledku formule pro nadpisy, resp. podnadpisy je vhodné nastavit v širokém rozsahu počtu slov: ** /T=50 **, resp. * /T=50 *).

** /T=20 /EPREUVE /T=20 **

Tento řetězec zúží výsledky na titulky obsahující jakékoli sloveso a vyzdvihuje titulky obsahující otazník nebo vykřičník (ve zpravodajství neobvyklé – mohou naznačit polemiku). Další řádky potom roztřídí výsledky podle vybraných kategorií sloves – jde o:

** /T=20 /CF=SpeechActs /T=20 ** (seskupení mluvních aktů jako asertivy, direktivy, komisivy, akreditivy, deklarativy a expresivy);

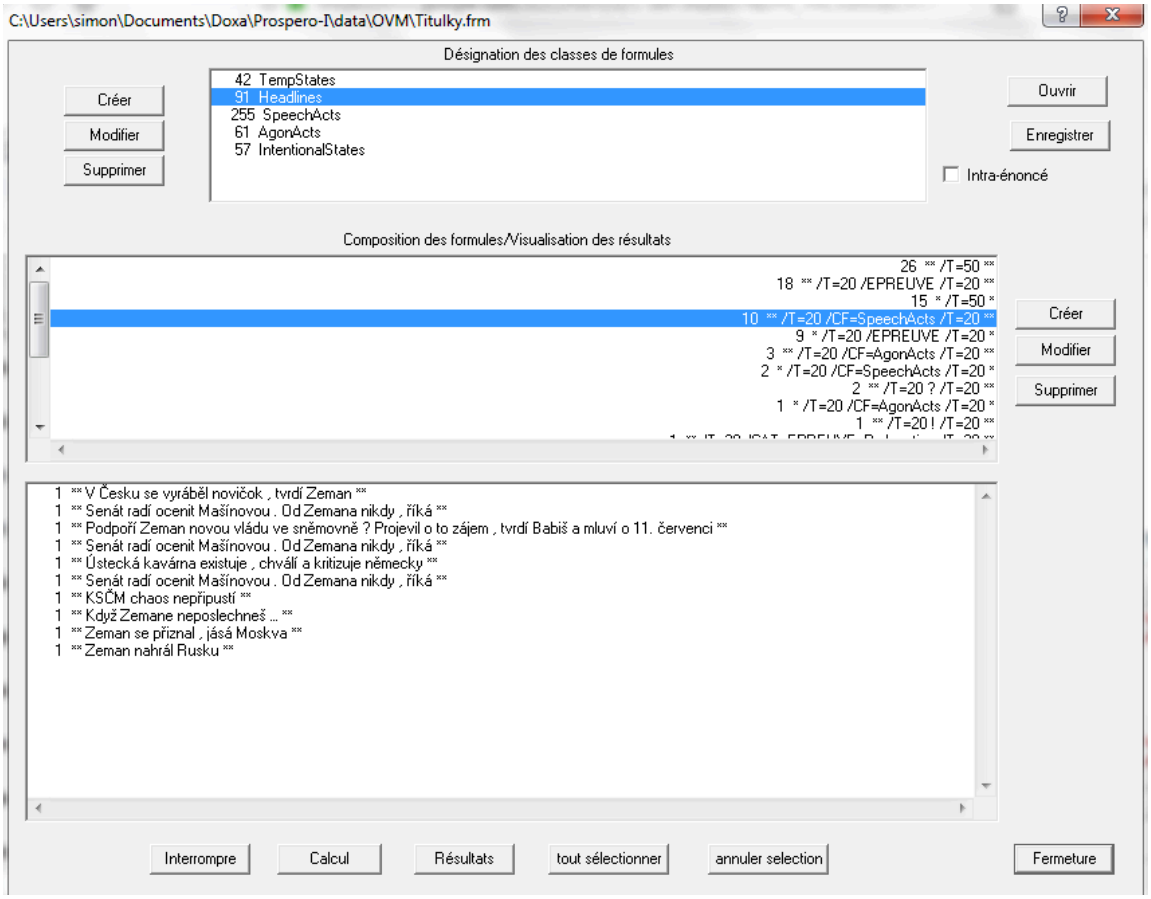
** /T=20 /CF=IntentionalStates /T=20 ** (intencionálních stavů jako myšlení, vědění, vnímání atd.);

** /T=20 /CF=AgonActs /T=20 ** (agonistických aktů - souhlas a nesouhlas);

** /T=20 /CF=TempStates /T=20 ** (časových stavů - pokračování, přerušení, předcházení).

Tyto řádky používají logický operátor /CF, který umožňuje seskupovat příkazy podobného typu (tady např. příkazy vyhledat sémanticky nebo syntakticky blízké kategorie sloves). Nejprve vytvoříme (pojmenujeme) kategorie formule v horním okně a kliknutím na název vytvoříme všechny náležité příkazy v dolním okně. Potom můžeme odkázat na název kategorie formule (ve formátu: /CF=název) v hlavní zóně naše formule (tady pod hlavičkou Headlines).

Obrázek níže ukazuje 10 výsledků pro titulky obsahující mluvní akty. Proč se objevuje jeden titulek třikrát? Protože obsahuje hned tři mluvní akty – ***radí*** (jde o direktiv), ***ocenit*** (expresiv) a ***říká*** (asertiv).



Přidáme-li do formule další řádky podle našich specifických potřeb a výzkumných otázek, můžeme tento prostor využít k rychlému zkoumání „zkrácené“ podoby korpusu, kterou představují jen titulky článků. Jde o „vertikální“ způsob rozdělení korpusu, které není možné použitím funkce podkorpusu, která rozděluje korpus „horizontálně“ (zejména tematicky) – za základní jednotky dělení bere celé texty.

Další formule v instalačním balíku Prospéra popisuje příloha 2.

Práce s formulemi vyžaduje určitou trpělivost a metodu "pokus - omyl", resp. zpřesňování dotazu, kdy se výzkumník pohybuje při procházení výstupů mezi dvěma extrémy - jednak se občas na řádku výsledků neukáže vůbec nic - jindy je výstupů tolik, že je nelze hned smysluplně interpretovat (je příliš mnoho „šumu“, a je třeba vyhledávací dotaz dále vymezit). Neexistuje jeden správný postup (odvíjí se nakonec od výzkumného záměru).

Do toho všeho ale vstupuje alespoň minimální znalost korpusu (jak byl konstruován) + přecházení mezi formulemi a konkrétními větami/texty. Dobrou metodou je přepsat známé spojení do formule. Např. víme, že tam určité "vzorové spojení" je, ale spojení nenacházíme - program tak musíme "naučit" dívat se určitým směrem a způsobem.

IV. Přílohy

Příloha 1: Definice kategorií

Entités

Absence of crisis	terms for calm periods
Actants	actors and roles – anything that can fulfil the role of agent, patient or receiver
Agir Stratégique Machiavélique	terms for scheming and duplicity
Argumentation	entities that refer to arguments and their protagonists
Authority-expertise	reference to experts and expertise, generally in an authority argument
Career	terms used to describe professional and political careers
Commitment	adoption/expression of a 'wanting to do'
Communication	entities involved in communication/mediation
Competence	adoption/expression of a 'being able to do'
Corruption	terms for corruption and corrupt actions
Crises et catastophes	terms for crisis and catastrophe
Democracy	entities that refer to a democratic topos
Dénonciation	terms used to make denunciations
Déploiement des conséquences	terms for describing causes and consequences
Déterminisme et Nécessité	terms that refer to historical determinism or fate
Discours sécuritaire	securitarian discourse
Economics	entities that belong to the economic/financial field

Elements-de-politique-publique	entities that belong to the public policy field
Emotion	entities that refer to an emotional/passional register
Emplois et conditions de travail	entities that belong to the sphere of work and employment
Ethique et morale	terms for ethical and moral principles or systems
EU	entities referring to the European Union
Failure	terms for failure
Formes juridiques	entities that belong to the legal field
Formes de responsibilité	terms for attributing responsibility for success and failure
Games	terms for games and game-making
Gestion des risques	risk management discourse
Incompetence	expression of a 'being unable to do'
Infrastructure	entities that refer to infrastructure and systems
Inquiétude	terms for describing fears and anxieties
Institutional reality	entities that exist because of institutionalising processes
Investigation	entities that refer to investigating and proving
Knowledge	entities that refer to the production of knowledge
Logique d'alarme	terms used to give alarms, alerts, warnings
Logique de décision et d'action	terms for describing decisions and actions
Manipulation	transfer/expression of a 'having to do'
Métalangage	terms for linguistic entities – talk about talk
Metalanguage of vilification	terms for vilification – talk about abusive talk
Modes de protestation	terms for describing protests and their protagonists
Narrative moments	the progression of a story, narrative phases
Narrativity	terms for stories and storytelling

Negotiation-consensus	entities that refer to the negotiation of consensus and compromise
Performance	narrative action and outcome
Personal relations	ties between actors, especially familial
Politics	entities that belong to the political field; political actions
Prospective	terms used for describing the future
Public mandate	publics in whose name actors claim speech rights
Raisonnement statistique	terms used in statistical reasoning
Régime de controverse	entities that refer to controversies and their protagonists
Rhétorique de changement	terms for describing or calling for change
Rhetoric of stability	terms for describing or calling for stability
Rhetoric of stagnation	terms for describing and denouncing stagnation / lack of change
Sanction	transfer/expression of an object of sanction
Secret-confidentialité	terms used to indicate concealment of information
Sociologie politique	terms that belong to the lexicon of political sociology
Status functions	functions that are conferred and carry entitlements or obligations
Temporalité	terms that refer to temporal progression/distribution
Witness	entities that refer to first-hand knowledge of events

Qualités

Accusation-Critique	qualities deployed to accuse/critique
A little	a small quantity or a lack of …
A lot	a large quantity of …
Ancien	old things
Apologetics	qualities deployed to approve

Assurance-Certitude-Fiabilité	stable/reliable things
Commitment	qualities that say something is 'wanted'
Compatible	qualities that insert something in a certain schema (-telný)
Dangerosité	dangerous/risky things
Emotional	qualities that attribute emotional dispositions to actants
Ideological	ideological qualities
Illegitimate	qualities that confer illegitimacy (but not necessarily accusatory)
Important-Essentiel	qualities that attribute importance to things
Incompatible	qualities that say something cannot be inserted in a certain schema (ne-telný)
Informational value	qualities that gauge the value or credibility of information
Institutional	qualities that imply institutionalisation
Legitimate	qualities that confer legitimacy
Likely	qualities that say something is likely
Manipulation	qualities that say something is obligatory
Moral	conformity of things to moral standards
Non-standard	non-standard/abnormal things
Nouveauté	new things
Politique	political things
Precedence	qualities that insert something in historical sequences
Standard	standard/normal things
Temporal	qualities that attach temporal values to things
Terminus technicus	qualities that belong to specialist fields
Unimportant	qualities that attribute a lack of importance to things
Unlikely	qualities that say something is unlikely

Marqueurs

Actualité	references to the present
Affaiblissement	weakening modalisation (markers of weak speaker commitment)
Anonymity	phrases to introduce an anonymous source
Approbation	markers of approval
Au-nom-de	ways to claim authority 'in the name of' a principle (principal)
Autonymie	phrases to designate linguistic objects, pause over the meaning of words
Biographical time	phrases that locate events with reference to biographical time (of speaker)
Calendrier	dates (months, days)
Causality	connectors that emphasise causality
Commitment	markers of the will to act
Comparateurs	phrases used to make comparisons
Competence	markers of the capacity/ability to act
Connecteurs de conséquence	consequential connectors (B follows from A)
Connecteurs de contradiction	contradictory connectors (B in spite of A)
Contextualisation	markers that specify the context of an assertion
Contradiction-réfutation	ways of refuting an assertion
Critique subtile-ironie-insinuation	phrases that convey mild criticism, irony or insinuation
Dedramatisation	ways to diffuse the drama of a situation
Definition of situation	ways of saying 'what it's about', 'what's at stake'
Dénonciation	markers used for denouncing
Dévoilement	connectors used to reveal what something really means
Doute-incertitude	markers of doubt

Dramatisation	ways to add drama to a situation
Durée	phrases used to mark lasting states
Embrayeurs	deictic expressions (pointing to the time, place and situation of enunciation)
Exemplification	ways to cite examples
Eventualité	conditionality ('if' words)
Evidence	marking an assertion as self-evident or obvious
Explicitation	making explicit or 'concrete' what something means
Failure	markers of failure
Fatality	markers of fatality
Généralisation	markers of generalisation
Greetings	markers of introductory and closing remarks in dialogue
Implication-Prise en charge	markers of explicit speaker commitment
Incompetence	markers of the incapacity/inability to act
Inflexion	ways of saying 'but'
Interdiscursive-markers	interdiscursive markers at the level of the text, references to interlocutor
Intuition-improvisation	markers of intuitive or improvised reasoning
Institutionally	markers of institutionality
Irréduction	markers that reject semantic reduction
Irreversibilité	markers of irreversibility
Journalistic distance	phrases used, typically in journalism, to attribute a proposition to a defined or undefined source
Le-Pire-Est-Devant-Nous	markers of doom
Lists and dates	notation for giving lists and dates
Manipulation	markers of the obligation to act
Négation	ways of negating an assertion
Normativity	markers referring to norms and rules

Orientation vers le futur	references to the future
Orientation vers le passé	references to the past
Paratextual marks	significant non-linguistic marks
Path dependency	markers of predestination
Point de basculement	markers of turning points
Political time	phrases that refer to political time
Polyphony	markers of polyphony (echo of a voice other than the speaker's)
Positionnement temporel	phrases used to place events in time
Positions-alliances	phrases that position actors as allies, supporters, helpers, etc.
Postponement	ways of marking the postponement of an event
Précaution rhétorique	markers of concession
Précédent	phrases used to invoke precedents or mark an event as unprecedented
Predictions	markers to introduce predictions or spekulations
Signalétique narrative	markers of narrativity
Proverb-metaphor	proverbial and metaphorical markers
Publicity	phrases that locate events in the public domain
Questions	common questioning introductors
Réalisme	markers that gauge what is realistic
Réduction	markers of semantic reduction
Renforcement	strengthening modalisation (markers of strong speaker commitment)
Répétition	phrases used to mark repeating states
Sanction	Phrases used to attribute credit and blame
Speed	phrases used to mark the speed of events
Statistique	markers of statistical reasoning

Status	phrases that qualify an assertion by reference to the status of the speaker
Sur le net	references to the Internet as a source or setting
Summary-simplification	markers to introduce a summarising or simplifying assertion
Témoignage	phrases that qualify an assertion by reference to the first-hand experience of the speaker
Test of time	ways of saying 'time will tell'
Time of crisis	phrases that locate events in a time of crisis
Time of stability	phrases that locate events in a time of stability
Urgence	ways to convey urgency
Warning	phrases that introduce warnings

Epreuves

Abandon	walking out, walking away, resigning
Accreditive	speech act, world-to-word direction of fit (Searle)
Agreement	reacting to a point of view by affiliation
Assertive-informative	speech act, word-to-world direction of fit (Searle)
Belief	intentional state, mind-to-world direction of fit (Searle)
Commissive	speech act, world-to-word direction of fit (Searle)
Commitment	'wanting to do'
Communicate	communicating, responding (degree zero)
Competence	'being able to do', 'knowing how to do'
Competition-Contest	agonistic verbs
Continuity	stable state, lasting
Declarative	speech act, double direction of fit (Searle)
Directive	speech act, world-to-word direction of fit (Searle)
Disagreement	reacting to a point of view by disaffiliation
Discontinuity	changes in state

Discuter	discussing, debating, arguing
Dissimuler-Mentir	deceiving, not telling the truth
Echouer	failing, not working (out)
Expressive	speech act, double/null direction of fit (Searle)
Happening	taking place, occurring
Ignorer-Oublier	not knowing, not understanding
Imagination	intentional state, null direction of fit (Searle)
Incompetence	not 'being able to do', not 'knowing how to do'
Knowing-Perception	intentional state, mind-to-world direction of fit (Searle)
Manipulation	'having to do'
Memory	intentional state, mind-to-world direction of fit (Searle)
Narrate	storytelling
Performance	doing
Perlocution	verb describing actual effect of speech act
Predict-precede	expecting, foreseeing
Prove-Define-Calculate	proving, defining, calculating
Representation	representing, standing for something
Sanction	giving/receiving an object of sanction
Thought	intentional state, mind-to-world direction of fit (Searle)
Wish	intentional state, world-to-mind direction of fit (Searle)

Příloha 2: Formule v instalačním balíku

Název	**Popis**
Variations on X	Zkoumá/mapuje/prochází diskurzivní okolí jevu X, kterým může být libovolné slovo, koncept, aktér nebo jev. X se definuje na řádek zvlášť, je to samostatná formule, která se vkládá do další formule. Formule je užitečná pro úvodní zkoumání pojmu nebo aktéra, ale i pro složitější skládání úrovní dotazů vyhledávání.
Citations	Hledá citace v různých variacích. Třídí výsledky podle typů použitých sloves.
RekIX	Hledá autory citací v běžném tvaru: "...", řekl X. Třídí výsledky podle typů použitých sloves.
Titulky	Generuje seznamy všech titulků a mezititulků v korpusu. Jedna variace formule zúží výsledky na titulky obsahující jakékoli sloveso. Dále třídí výsledky podle vybraných kategorií sloves – jde o seskupení mluvních aktů (asertivy, direktivy, komisivy, akreditivy, deklarativy a expresivy), intencionálních stavů (myšlení, vědění, vnímání atd.), agonistických aktů (souhlas a nesouhlas) a časových stavů (pokračování, přerušení, předcházení). Další variace formule vyzdvihuje titulky obsahující otazník nebo vykřičník. (Pozn.: je funkční pouze za předpokladu, že titulky jsou označeny dvěma hvězdičkami na začátku a na konci a mezititulky jednou hvězdičkou na začátku a na konci!)
Matters of concern	Hledá výrazy odkazující na pro mluvčího důležité věci, často vyslovené v kontextu argumentů autority, kde nějaký princip slouží jako zdůvodnění postoje, rozhodnutí nebo strategie (co je důležité pro koho, ve jménu čeho musíme/chceme konat atp.)