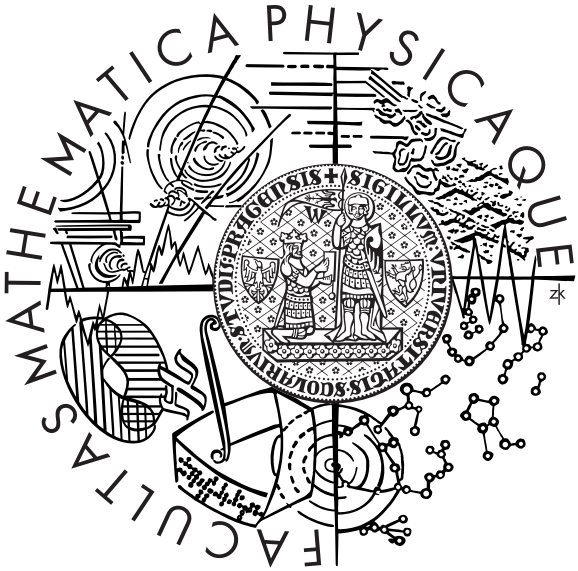
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jan Sochna

Aplikace pro ruční word alignment

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Pavel Pecina

Studijní program: informatika, správa počítačových systémů

2007

Děkuji Mgr. Pavlu Pecinovi za trpělivé vedení práce, poskytnutí části česko-anglického korpusu, na které bylo možné program vyvíjet, za jeho inspirativní nápady a pomoc při zakončování práce. Dále mé rodině, spolužákům a kamarádům za pomoc při testování uživatelské přívětivosti a podporu při práci.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 7. srpna 2007 Jan Sochna

Obsah:

Název práce: *Aplikace pro ruční word alignment*
Autor: *Jan Sochna*
Katedra: *Ústav formální a aplikované lingvistiky*
Vedoucí bakalářské práce: *Mgr. Pavel Pecina*
Email vedoucího: *pecina@ufal.mff.cuni.cz*
Abstrakt: *Cílem této práce bylo navrhnou a implementovat na platformě nezávislé, rychlé, flexibilní a přívětivé uživatelské rozhraní pro ruční párovaní (alignment) dvoujazyčných textů. Nové rozhraní nemá nedostatky existujících nástrojů na párování a proces ručního párování zefektivňuje. Jde např. o poloautomatické párování jednoduchých vět, skupinové operace s párováním, párování frází, možnost posunu jedné z párovaných vět podél řádku vůči druhé větě pro zlepšení přehlednosti, mají-li párované věty různou délku, přehledné zobrazování předchozího a navazujícího kontextu párovaných vět v obou jazycích a v neposlední řadě i statistika postupu párování. Vedle obvyklého řádkového pohledu - zobrazení párovaných textů ve dvou řádcích nad sebou, kdy se páruje propojením odpovídajících si slov čarou, byl realizován i pohled maticový - kdy slova věty v jednom jazyce odpovídají popisu řádků matice, slova v druhém jazyce odpovídají popisu sloupců matice a páruje se zvýrazněním průsečíku sloupce a řádku, které mají odpovídající si popisy. Mezi oběma pohledy lze během práce libovolně přepínat.*

Klíčová slova: *párování dvojjazyčných textů, word alignment, návrh GUI, maticové zobrazení*

Title: *Application for manual word alignment*
Author: *Jan Sochna*
Department: *Institute of Formal and Applied Linguistics*
Supervisor: *Mgr. Pavel Pecina*
Supervisor's e-mail address: *pecina@ufal.mff.cuni.cz*
Abstract: *The aim of this work was to design and implement platform-independent fast, flexible and user friendly interface for manual word alignment of bilingual texts. The new interface does not have the imperfections of existing similar tools and improves the performance of manual alignment process. It provides eg. half automatic alignment of simple texts, group operations with alignments, alignment of phrases, enables to shift one sentences along the line to improve the transparency of the alignment process in case that the length of aligned sentences differs substantially. The preceding and succeeding context of currently aligned sentences is shown in both the languages. Last but not least the tool provides the alignment performance statistics. Along with usual "row view", where the two sentences are shown in parallel in two rows, one above the other, being aligned by connections of corresponding words, there were introduced also a "matrix view", where the words in one language stand in for matrix line descriptors, the words in other language stand in for column descriptors and the alignment of two corresponding words is expressed by highlighting of the point of intersection of row and column with corresponding descriptors. It is possible to switch between the both views anytime during the alignment process.*

Keywords: *alignment of bilingual text, word alignment, GUI design, matrix view*

1 Úvod

Při word alignmentu (párování slov) se slovům z věty v jednom jazyce přiřazují slova z věty v jazyce druhém tak, aby se získaly nejlepší překladové páry slov. Párování slov se pak používá trénování strojových překladových systémů. Pro takové učení je třeba velké množství párovaných vět, které se získávají automatickým párováním slov. Výsledky ručního word alignmentu se používají právě pro učení těchto automatických párovacích nástrojů a pro vyhodnocování jejich úspěšnosti.

Další použití párování je při automatickém hledání správného významu slov ve volném textu *(word sense disambiguation*). Právě využití dvojjazyčných textů ke zvýšení přesnosti výsledků je označováno za inovativní [1].

V této práci se zabýváme návrhem a implementací rychlého a flexibilního rozhraní pro párování slov, přenositelného mezi počítačovými platformami. Důraz je kladen na odstranění nedostatků existujících aplikací a zrychlení procesu párování.

2 Vyhodnocení stávajících aplikací

Cílem bylo seznámit se s obdobnými aplikacemi a posouzení jejich kladných a záporných vlastností. Podrobněji jsem se seznámil s následujícími aplikacemi:

Cairo [2] – napsaný v jazyce Java.

Align [3] – napsaný v jazyce Java.

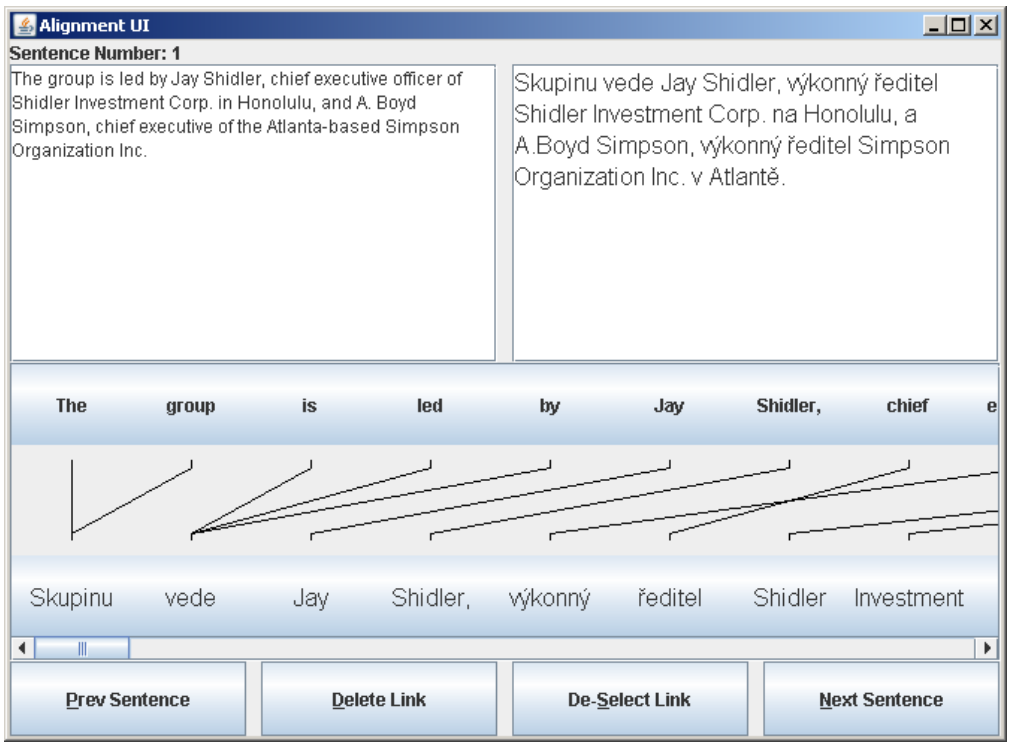
Alpaco [4] – napsaný v jazyce Perl.

Cairo, jako součást balíčku aplikací The EGYPT Statistical Machine Translation Toolkit, se zaměřuje na vizualizaci word alignmentu ale neumožňuje párování vytvářet. Proto se jím dále nezabýváme.

Nyní rozeberme postupně přednosti a zápory zbylých dvou aplikací.

2.1 Align

Obsahuje částečnou podporu pro volbu znakové sady. Umožňuje volbu kódování pro jednu jazykovou verzi textu. Pro text ve druhém jazyce již používá přednastavenou znakovou sadu.



Obrázek 1 - UMIACS Word Alignment Interface (zešikmené zobrazení)

Obrázek 1 ukazuje vzhled aplikace. Lze si všimnout problému s posunem vět, kdy se spojnice naklání a odpovídající si slova se od sebe vzdalují. Pro vytvoření spojení je pak zapotřebí většího úsilí.

Dalším rušivým prvkem je použití různých fontů pro věty a slova v různých jazycích. To může být výhodné pro texty v exotických jazycích jako je například čínština nebo japonština, ale u textů psaných latinkou působí rušivě.

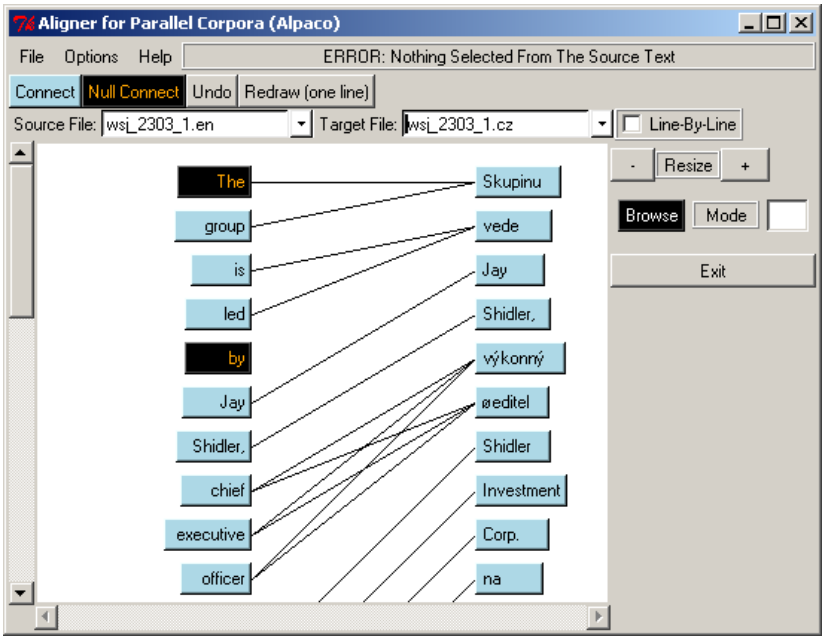
Mazání spojnic probíhá ve dvou krocích. Nejprve je třeba spoj vybrat a až následně je možné spojnici smazat (odpovídá 3 kliknutím).

V aplikaci nelze vytvářet prázdná spojení (*Null connection*) pro slova, která ekvivalent v druhé větě nemají. Uživatel je také naváděn k úplnému párování zadaných vět, což není vždy žádoucí.

Lze využít k párování jen jednoho bloku vět.

2.2 Alpaco

Už samotná instalace se ukázala být problematickou, neboť stažený soubor obsahoval chyby, kvůli kterým jej nebylo možné po instalaci spustit. Tento problém se projevil na různých hardwarových i softwarových platformách, i při opakovaném stažení aplikace. takže pravděpodobně nešlo o chybu jedné instalace. Obsažené chyby se podařilo odstranit a tak můžeme zhodnotit aplikaci samotnou. Vzhled aplikace ukazuje Obrázek 2



Obrázek 2 – Alpaco

Nevýhodou aplikace je, že nepodporuje jakoukoliv volbu vstupního kódování. Diky tomu se špatně zobrazují slova ve sloupci s češtinou. Předchozí použití konverzního programu by mohlo problém řešit, ale nemusí pomoci. Příčinou může být i nevhodně nastavené písmo.

U delších vět se opět objevuje problém s jednotným posuvem v obou jazykových verzích. Slibná jsou tlačítka pro změnu velikosti, avšak ovlivňují pouze mezeru mezi jednotlivými slovy v rámci zobrazení vět. Určitou výhodou je možnost přepínání mezi prohlížecím a editačním módem.

Zajímavá je možnost vrátit zpět poslední akci, avšak pokud poslední akce znamenala vytvoření většího počtu spojení, odebírání nefunguje hromadně, ale pouze po jednotlivých spojích. *Opětovné provedení akce* je následně také možné, a to i v případě, že od odvolání akce byla odvedena nějaká práce. Obecně odvolávání funguje jen pro odebírání spojů. Obnovení spojů po smazání je možné *opakování akce*.

Rušení spojení lze provádět několika způsoby. Přímým výběrem jedné spojnice označením konců a použitím undo. V takovém případě je odstraněno právě jedno spojení. Výběrem slova z věty v jednom jazyce a použitím undo. Zruší se všechna spojení vycházející z daného slova. Poslední možností je použití již dříve zmíněného undo bez jakéhokoliv výběru. Pak dojde k odstranění jedné z posledně vytvořených spojnic.

Lze vytvořit na slově *Null connection* a zároveň na něm zachovat obyčejné spojení. Což si logicky odporuje (na obrázku 2 je to vidět na prvním slově anglické věty „The“).

Modifikovaná verze Alpaca [4], dostupná odděleně od původního projektu přidává důvěryhodnost spojení pomocí značení *sure* a *possible*. To by mělo pomoci k získávaných přesnějších výsledků v navazujících projektech. Má však problémy při opětném otevírání souborů s uloženým párováním. Nevytváří při ukládání identifikační hlavičku souboru a zvolený soubor je při následném otevírání odmítnut.

Lze použít k párování i více bloků textu, avšak soubory je třeba vždy ručně nastavit do aplikace. Přechod mezi bloky si musí zajistit uživatel sám načtením dalších souborů.

2.3 Shrnutí

Společným problémem obou uvedených aplikací (Align, Alpaco) je volba kódování vstupních souborů a souběžný posun zobrazených vět.

Párované dvojjazyčné věty nejsou vždy stejně dlouhé. Z toho důvodu se spojnice odpovídajících si slov naklání a prodlužují. Přehlednost párování tak klesá, jak ukazuje obrázek 1-2.

Obě vyjmenovaná řešení mají také problém s mazáním spojnic. Je potřeba vždy vybrat spojnici kliknutím na její koncová slova a pak teprve lze stiskem klávesy či tlačítka na formuláři spojnici odstranit. Alpaco v tomto směru nabízí více možností.

Další hodnocenou stránkou programu bylo, jak aplikace využívá plochu, kterou má její okno k dispozici. Při maximalizovaném okně by měl uživatel získat větší komfort pro práci s aplikací. Lépe je na tom Align, který při zvětšování rovnoměrně roste. Zatímco v případě Alpaca se okno zvětšuje jen ve vertikálním směru a jinak se pouze zvětšuje nevyužitá plocha aplikace.

3 Návrh aplikace

Pro zajištění přenositelnosti bylo zvoleno programování v jazyce Java, který je úspěšně portován na operačních systémech Linux, Solaris a Windows. Do své platformy Mac OS X zařadil podporu Javy také Apple.

Cílem návrhu bylo vytvořit aplikaci, která by umožňovala párování v různých pohledech, a umožňovala tak využívat jejich specifických výhod. Proto byla aplikace rozdělena na dvě základní časti - jádro aplikace a grafické rozhraní. Popis jádra aplikace a způsob implementace pohledů je součástí programátorské dokumentace na přiloženém CD.

Grafické rozhraní bylo původně oddělené a každý pohled měl svůj vlastní formulář pro komunikaci s uživatelem. Tento návrh ovšem nevyhovoval, protože mnoho funkcí je společných všem náhledům. Proto je ve finální verzi grafické rozhraní rozděleno na dvě hlavní časti. Jedna obsahuje společné funkce v menu spolu s panelem nástrojů pro ovládání hlavních operací. Druhá obsahuje podokno pro samostatnou vizualizaci párování (pohled) a jeho ovládání.

Jelikož každý pohled na párování má svá specifika, vznikl v pozdější době návrhu model, kdy každý pohled může mít navíc svůj vlastní panel nástrojů, zpřístupňující specifické funkce pohledu a vlastní klávesové zkratky. Při přepnutí do jiného pohledu je původní panel odstraněn a nahrazen panelem pro nový pohled.

Aplikace je navržena tak, aby bylo možné přepínat mezi pohledy kdykoliv (i před uložením dat).

Součástí návrhu je také možnost pro omezení času stráveného na větě. Myšlenkou této funkce je fakt, že pokud uživatel příliš váhá, bude lepší, když bude pracovat na další větě, aby práce ubývala. Jak dlouhý tento časový limit je, či zda se vůbec bude uplatňovat lze upravit v nastavení programu.

Program během používání také generuje statistiky odvedené práce, využívání funkcí a času stráveného u aplikace, které je možné využít k následné analýze. Časem stráveným u aplikace se rozumí doba, kdy uživatel s programem skutečně pracoval a nikoliv doba spuštění programu.

3.1 Formát vstupu a výstupu

Vstupem programu jsou dvojjazyčné texty párované na úrovni vět (*sentence aligned*). Jedná se o dva seznamy souborů, kde každý seznam reprezentuje texty v jednom jazyce (bloky vět).. Dílčí soubor textů příslušného jazyka (odpovídající jednomu bloku) pak obsahuje na každém řádku jednu větu k párování. Pro správnou funkci programu si seznamy musí odpovídat počtem bloků i počtem vět v jednotlivých blocích

Pro vstup dvojjazyčných textů do úlohy lze využít i věty označené značkami (*tags)*. Je možné načítat různé formáty značek a nastavit masku značky do konfiguračního souboru. V případě, že je vstup označen za značkovaný, ale formát není uveden, hledají se na začátku vět značky ve tvaru:

```
<s id="blok:cisloVety">
```

nebo

```
<s id="jineJednoznacneOznaceni">
```

Tato identifikace se pak zobrazuje uživateli, aby měl představu, kde se při párování nachází. Vzhledem k požadavku na jedinečnost této autoidentifikace ji pak lze použít pro rozlišení vět i uvnitř programu.

Pro formát výstupu byla s výhodou využita stávající struktura z projektu Blinker [6]. Ta je dobře podporována ze strany word alignment nástrojů i programů, které párování využívají. Výstupem párování jedné věty je soubor ve jmenné konvenci:

```
samp<cisloBloku>.SentPair<indexVety>
```

V každém ze souborů je dvojice sloupců oddělená mezerou. Každý řádek souboru pak symbolizuje jedno spojení mezi slovy. Obsahuje indexy odpovídajících si (spárovaných) slov oddělené mezerou (výstupním separátorem). Pokud je nějaké slovo označené jako *Null connection*, tj. že v textu ve druhém jazyce nemá žádný ekvivalent, má ve sloupci druhého jazyka uloženu nulu. Indexy regulérních spojení začínají od jedničky.

Volitelnou možností je doplnit do třetího sloupce výstupu textově také odpovídající si slova, pro možnost vizuální kontroly textových výstupů. Tato funkce je rozšířením nad rámec definice výstupního formátu Blinkeru.

Část výstupního souboru pak může vypadat například takto:

```
0 17 (v)
2 1 (consortium, Konsorcium)
4 2 (private, soukromých)
5 3 (investors, investorů)
6 4 (operating, fungující)
7 5 (as, jako)
```

Výstupní statistiky generované programem mají tvar podobný souboru s nastavením viz. níže. Statistikou je soubor s názvem „stat_YYYYMMDD-HHMMSS.log“, kde velká písmena označují po řadě datum a čas začátku párování. Soubor pak obsahuje jednotlivé záznamy jako například:

```
Paired-Sentences=8
Created-Connections=124
Deleted-Connections=20
Time-Spent-On-Aplication=11
Time-Exceed=3
```

Tento zápis vykazuje počty pro: zpracované věty, vytvořená a zrušená spojení, čas strávený u aplikace v minutách a počet překročení časového limitu na větu za dobu práce.

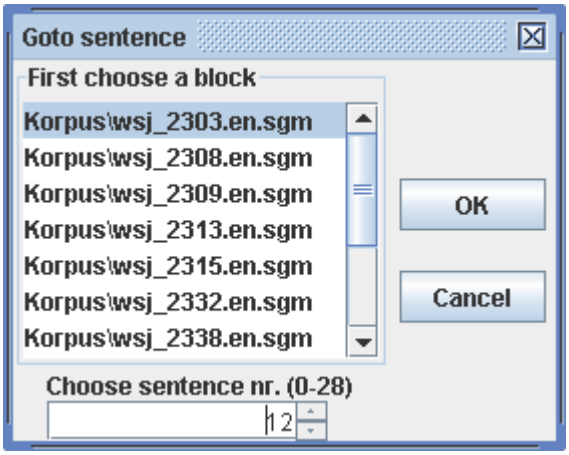
3.2 Společné části grafického rozhraní

V záhlaví grafického rozhraní je menu a panel nástrojů (k vidění na obrázcích 3, 7 a 11), který obsahuje základní funkce aplikace. Přechod na další nebo předchozí větu lze učinit pomocí stisku příslušného tlačítka, nebo pomocí klávesy „N“ (Next) pro další větu, „P“ (Previous) pro předchozí větu.

Obrázek 3 - Detail panelu nástrojů

Další způsob přechodu na jinou větu je zadáním cílového bloku a čísla věty v tomto bloku. Dialogové okno pro tuto operaci je přístupné přes menu *Control* > *Goto sentence* ... nebo pomocí klávesové zkratky CTRL+G. Náhled dialogového okna lze vidět na obrázku 4.

Při žádném z těchto přechodů se změny neukládají. Pro uložení je třeba použít tlačítko s vyobrazením diskety, nebo položku menu *Save*, nebo klávesovou zkratku CTRL + S. Po uložení aktuální věty pokračuje párování hned větou další.



Obrázek 4 - Goto sentence formulář

Dále je na panelu nástrojů umístěné ovládání historie operací: Undo a Redo. Příslušné ovládací prvky se aktivují, až když mohou nabídnout nějakou akci. Undo je stavěné tak, aby odvolávalo opravdu celou poslední operaci a nikoli jen její část. Odvolání části je sice v systému zabudováno, ale není zpřístupněno uživateli. Je to dáno tím, že ač algoritmem je pořadí vzniku či rušení spojnic pevně dáno, jeho vysvětlování by bylo zbytečně složité a celkově je taková možnost obtížně sdělitelná.

- **Undo** se chová podobně, jako můžeme být zvyklí z textových editorů. Tedy je možné odvolávat poslední akce až do původního stavu. Počet kroků, které lze odvolat není přímo omezen. Paměť pro Undo se však maže přechodem na jinou větu.
- **Redo** se chová také známým způsobem. Lze znovu provést poslední odvolané operace. Pokud se po odvolání operace provedla operace úplně nová, je zbytek *redo bufferu* smazán a tyto operace již nelze opakovat! Příslušná ovládací prvky jsou v tomto případě neaktivní.

Tlačítko pro triviální párování je posledním na této části panelu. Umožňuje vytvořit jedním stiskem spojení mezi stejně položenými slovy ve větě. Vzniknou tedy spojení vždy mezi i-tým slovem textu v jednom jazyce a i-tým slovem textu v jazyce druhém. Tuto funkci může uživatel využít v případě, kdy bude zpracovávat

dvojici vět, které budou mít mezi sebou zcela nebo z větší části takto jednoduchý vztah.

Další část panelu nástrojů umožňuje přepínání mezi zobrazeními. Vedle jednoduchého náčrtu, který zobrazení charakterizuje, je vždy ještě uveden textově název zobrazení. Aktuálně používané zobrazení je zvýrazněno podbarvením. Tuto část panelu lze nalézt se stejnou funkčností i v menu *View*. Jednotlivým pohledům jsou přiřazeny klávesové zkratky vycházející z funkčních kláves, které jsou na klávesnici odděleně. Prvnímu pohledu je přiřazena klávesa F1, dalšímu F2, kdyby bylo pohledů více mohly by mít přiřazeny další funkční klávesy.

Obrázek 5 - detail panelu pro výběr zobrazení

Obrázek 6 - detail panelu pro ovládání lupy

Ovládání velikosti zobrazení se nachází na dalším samostatném panelu nástrojů (obrázek 6). Lze jím ovlivnit čitelnost jednotlivých zobrazení. Ačkoliv výchozí nastavení velikosti by mělo být odvozeno od systémového nastavení, a tedy zajišťovat dostatečnou čitelnost, je možné je pomocí tlačítek lupy na hlavním okně po skocích měnit. Funkce zámku lupy (*zoom lock)* umožňuje zafixovat nastavení lupy tak, aby se zachovalo i při přepnutí mezi pohledy. O stavu funkce *zoom lock* je uživatel informován grafickým obrázkem zámku, který je buď odemčený nebo zamčený a symbolizuje tak fixaci či volnost aktuálně nastaveného zvětšení. Pokud je funkce *zoom lock* vypnutá, pak se nastavené zvětšení zachovává jen do nejbližší změny pohledu. Změnou pohledu není přechod na jinou větu.

Vedle tohoto panelu je prostor pro panel nástrojů aktuálního pohledu. Zde je možné zpřístupnit funkce specifické právě danému pohledu.

Pokud má na jednom slově existovat *Null connection* a obyčejné spojení, nazveme takový případ kolizí. Aplikace neumožňuje vznik kolize buď tak, že nedovolí vložit informaci způsobující kolizi nebo původní informaci nahradí nově vkládanou. Způsob reakce závisí na nastavení (viz. vlastnost *force-creation*).

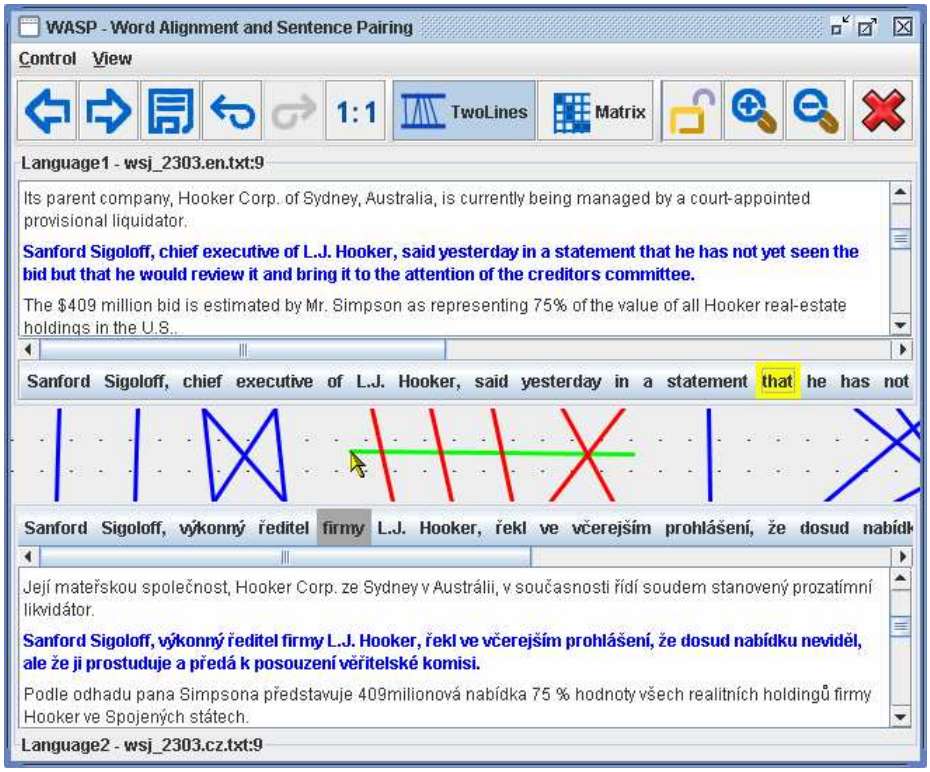
Druhá možnost je nastavena jako výchozí, protože odpovídá modelu, že uživatel ví dobře, co dělá a operaci si přeje provést. Pokud by nové spojení bylo umístěné chybně, pak ho může uživatel snadno vzít zpět pomocí *„undo“*.

4 Implementované pohledy

Podařilo se implementovat dva pohledy – dvouřádkový a maticový. Dvouřádkový vychází z existujícího řešení Align [3] a Alapaco [4]. Maticový pohled je pak realizován nově.

4.1 TwoLines view

Je prvním implementovaným pohledem. Dostupnou plochu rozděluje na tři části, z nichž okrajové části zobrazují okolní kontext obou vět. Kontext může být buď omezený na zadaný počet vět nebo neomezený, kdy se zobrazují všechny věty, které jsou dostupné. Původně byl kontext zobrazen jako jednoduchý text, ale přidáním formátování se výrazně zvýšila přehlednost. Aktuální věta je v rámci kontextu barevně zvýrazněna a snahou je, aby byla i vertikálně vycentrována a tedy byly čitelné i bezprostředně sousedící věty. (Obrázek 7)



Obrázek 7 - Ukázka rozhraní s řádkovým pohledem

Ve střední části pohledu se realizuje samotné párování. Slova aktuální věty se zobrazují na okraji plochy sloužící k vizualizaci párování. Pokud se zobrazená věta do okna nevejde, aktivuje se v části přilehlé ke kontextu posuvník.

Standardně se posuvníky u objektů zobrazují vpravo nebo dole. Zde umístění posuvníku dole vadilo u vizualizace věty v prvním jazyce (v horní časti formuláře), neboť odděloval slova od plochy pro vizualizaci spojnic. Úplné vypnutí posuvníků se také neukázalo jako ideální. V tomto případě sice plocha oddělena nebyla, ale uživatel ztratil přehled o délce věty a své pozici ve větě. Implementované řešení, s posuvníkem schovaným do zobrazení kontextu, plní funkci orientační, kdy uživatel vidí, jak daleko se nachází ve větě. Navíc ještě nezasahuje do prostoru vizualizace spojnic a tedy není rušivým elementem.

Každé slovo z párovaných vět je zobrazeno formou tlačítka, které mění barvu podle aktuální situace.

- Bez zvýraznění – výchozí stav slova
- Žlutá – vybrané slovo v daném jazyce
- Šedá – *Null connection* na slově
- Šedo žlutá – *Null connection* na vybraném slově

Jak již bylo zmíněno, plocha mezi větami v jednotlivých jazycích slouží pro zobrazení spojení odpovídajících si slov. Je rozdělena pomocí dvou přerušovaných čar na tři pásy, které pomáhají uživateli při ovládání posunu vět a spojnic slov, jak je popsáno dále.

K posunu v aktuální větě může uživatel využít buď běžné posuvníky umístěné v blízkosti vět nebo může vyžít funkci „ruky“, kdy jednoduše uchopí někde v oblasti párování plochu a odtáhne ji tak, aby mu zobrazení vyhovovalo. Uchopí-li plochu v jednom z krajních pásů, posouvá přilehlou větu, uchopí-li plochu ve středovém pásu, pohybuje oběma větami souběžně.

Spojnice je reprezentována čarou, která vychází ze středu odpovídajících si slov. Lze ji vytvořit výběrem slov z vět obou jazykových verzí, které se mají spojovat. Slovo se vybírá kliknutím levého tlačítka myši (na obrázku je vybraným slovem „that“). Jakmile je ve výběru alespoň jedno slovo z každé z párovaných vět, vzniká mezi vybranými slovy spojnice. Ta je ve výchozím stavu modré barvy; pouze pokud je spojnice vybrána, je barvy červené.

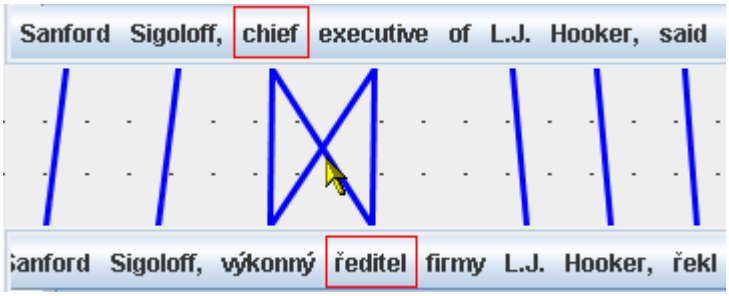
Výběr spojnic se provádí intuitivně a velmi snadno pomocí kliknutí přímo na spojnici (zde je drobná tolerance na vzdálenost od čáry), kdy se vyberou všechny

spoje pod kurzorem. Nad kterou spojnicí se uživatel nachází vidí podle zvýraznění spojených slov (obr. 8). Pokud je třeba vybrat větší množství spojnic, lze využít alternativní mód výběru. Stiskem pravého tlačítka mimo spojnice a jeho tažením po ploše lze vybrat všechny spojnice, které takto nakreslená čára protne (zelená čára na obrázku 7 reprezentuje vybírací spojnici a červené spoje vyznačují spoje, které takto nakreslenou čarou byly vybrány).

Výběr lze zrušit buď opětovným kliknutím na spojnici, nebo pomocí dvojitého stisknutí klávesy ESC (první stisk přeruší právě prováděnou operaci, pokud nějaká probíhá; druhý již zruší výběr slov)

S vybranými spojnicemi je možné dále pracovat. Je možné je přesouvat nebo hromadně smazat stiskem příslušného tlačítka na panelu nástrojů (nebo klávesou DELETE).

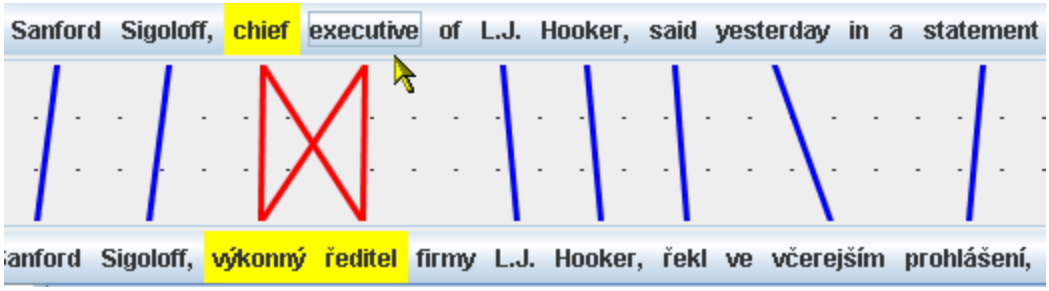
Mazání spojnic se jednotlivě provádí kliknutím pravým tlačítkem na spojnici. Nad jakou spojnicí se uživatel nachází, signalizuje rámeček kolem spojených slov (obr. 8). Uživatel má tedy dobrou kontrolu nad tím, jaké spojení se bude mazat.



Obrázek 8 - Zvýraznění spojených slov při přejetí myší přes spojnici

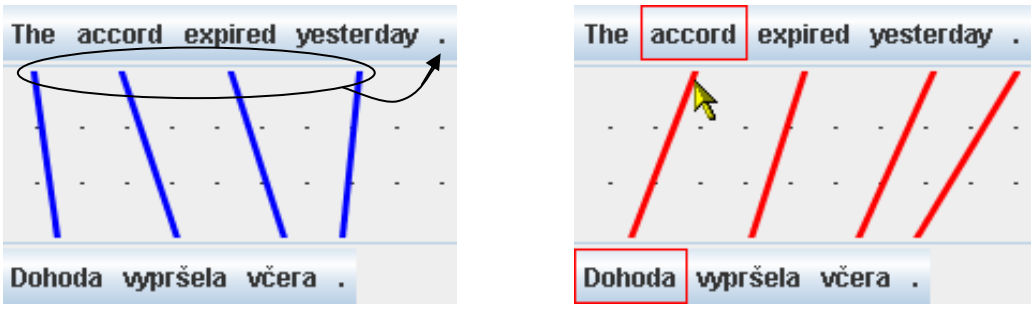
Null connection se vytváří kliknutím pravým nebo prostředním tlačítkem na slovo, které chce uživatel označit. Vizualizace je provedena změnou barvy pozadí slova na šedivou (na obrázku 7 se jedná o slovo „firmy“)

Pokud je třeba vytvořit frázové spojení (*phrasal connection*) na více slovech, pak je možné k vybírání podržet klávesu CTRL, tím se zabrání předčasnému vytvoření spojení. Až před výběrem posledního slova je třeba klávesu CTRL pustit. Následný výběr posledního slova spustí vytvoření příslušných spojnic. Tedy úplné bipartitní spojení.



Obrázek 9 - vytváření frázového spojení

V situaci, kdy si slovosled v obou větách odpovídá, lze využít funkci pro *triviální párování*. Vytvoří se spojnice vždy i-tého slova věty v jednom jazyce s i-tým slovem věty v druhém jazyce. V případě, kdy takto utvořené páry nevyhovují kvůli drobnému posuvu (například kvůli absenci členu nebo částice v jednom z jazyků), lze posunout konce všech vybraných spojení na straně textu v jednom z jazyků o shodný počet pozic tak, aby spojnice odpovídaly párovanému textu. viz. Obrázek 10.



Obrázek 10 – modifikace triviálního párování (před a po posunutí)

Spoje je třeba vybrat (viz výše) a uchopit v oblasti pásu přiléhajícím k větě, ve které chceme realizovat posun konců spojnic a přetáhnout je na nové místo. Nelze přesouvat konce spojnic v obou větách zároveň. Při přetahování se průběžně ukazuje, kam vybrané spojnice míří. Po uvolnění myši se přesunutá spojení zafixují. Operaci přesunu lze přerušit stisknutím klávesy ESC.

Zvýšení přehlednosti přispívá poměr v jakém se zobrazují jednotlivé části pohledu. Zvětšením okna aplikace se zobrazení kontextu zvětší relativně více, než plocha pro párování. Zatímco bohatší kontext představuje přínos pro uživatele, zvětšením plochy pro párování (vertikálně) by došlo jen k prodloužení spojnic.

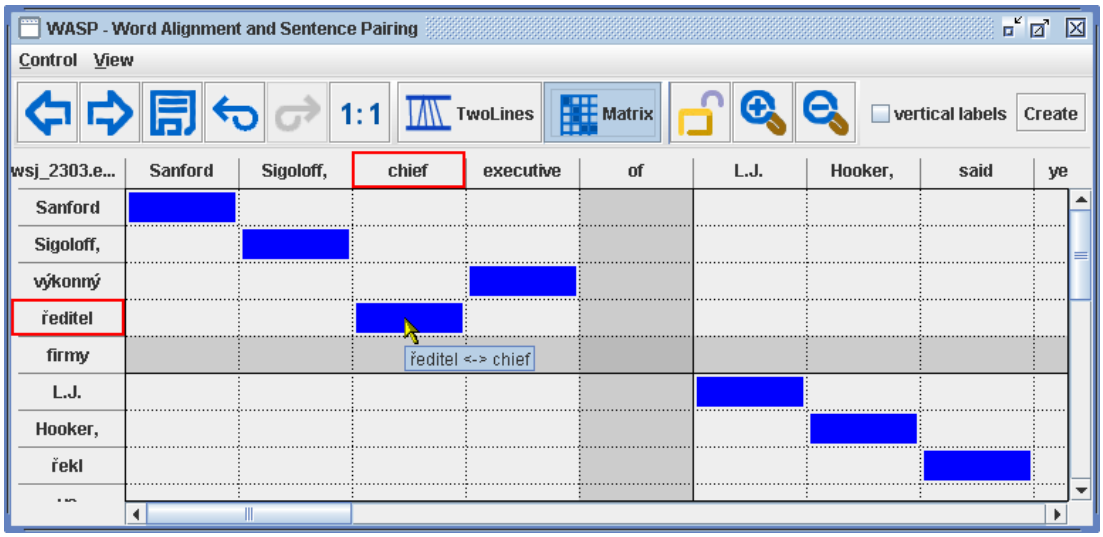
Dalším praktickým vylepšením je možnost omezit zobrazovanou délku slov. K čemu taková věc slouží? Některé jazyky jsou známé svými složeninami (například němčina, finština). Texty v těchto jazycích by při vizualizaci zabíraly příliš místa, pokud by se slova zobrazovala v celé jejich délce. Proto nastupuje omezení slov na maximální zobrazovanou délku, kdy se konec slova zkracuje do tří teček „…“. Původní znění se pak zobrazuje jako plovoucí nápověda.

Použití funkce lupy na tomto pohledu mění velikost použitého písma v kontextu a vizualizaci slov. Obdobně také ovlivňuje omezení na maximální šířku slov.

4.2 Matrix view

Tento pohled vznikl podle námětu vizualizace word alignmentu v literatuře, avšak s jistými změnami. Zobrazení není podobné mřížce či X-Y grafu, ale spíše tabulce (Obrázek 11). Hlavním přínosem této reprezentace je vetší kontrola nad umísťováním spojů a vytváření spojení jediným kliknutím. Metoda nulové tolerance, která vyžadovala, aby uživatel při vytváření spojů nepohnul kurzorem, byla sice nejsnazší na implementaci, ale příliš stěžovala vlastní práci s aplikací. Zavedením drobné tolerance lze spoje vytvářet mnohem svižněji. Tabulka má navíc zvýrazněné bloky 5x5 polí, aby se uživateli neslévala a snáze se orientoval. V levém horním rohu tabulky se zobrazuje název aktuálního bloku spolu s indexem aktuální věty pro orientaci v postupu mezi bloky a v rámci aktuálního bloku. Tuto identifikaci je možné uplatnit při použití funkce *Jump to sentence.*

V záhlaví řádků a sloupců se zobrazují slova vět jednoho resp. druhého jazyka. Obarvené pole na jejich průsečíku vyznačuje existující spojení mezi těmito slovy. Dle pozice kurzoru se zvýrazňují slova v příslušném řádku a sloupci. Současně se u kurzoru objevuje plovoucí popisek (*tooltip*) naznačující potenciální spojení zvýrazněných slov (ukazuje mj.Obrázek 11).



Obrázek 11 - Ukázka rozhraní s náhledem Matrix

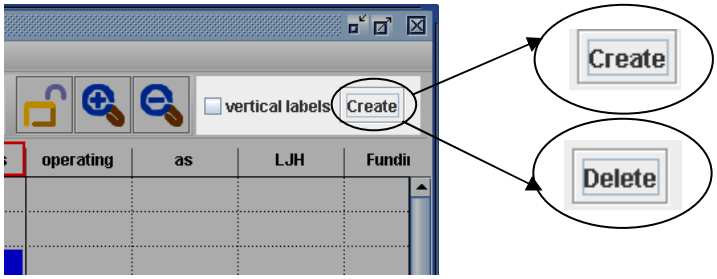
Pokud nějaké slovo nemá ve větě druhého jazyka ekvivalent a je třeba je označit pomocí *Null connection*, pak se tak provede kliknutím pravým tlačítkem na toto slovo. V tabulce se to projeví ztmavnutím příslušného řádku či sloupce (k vidění pro slova „firmy“ a „of“).

Posun v tabulce je opět vyřešen funkcí „ruky“. Tato operace může na pomalejších počítačích způsobovat „cukání“ obrazu z důvodu nároků na vykreslování tabulky.

Vytváření a rušení jednotlivých spojů je možné pomocí stisku pravého tlačítka myši nad požadovaným polem. Vytváření frázových spojení je omezeno na vytváření spojení mezi skupinami sousedících slov (vyznačeno jako souvislá plocha - obdélník).

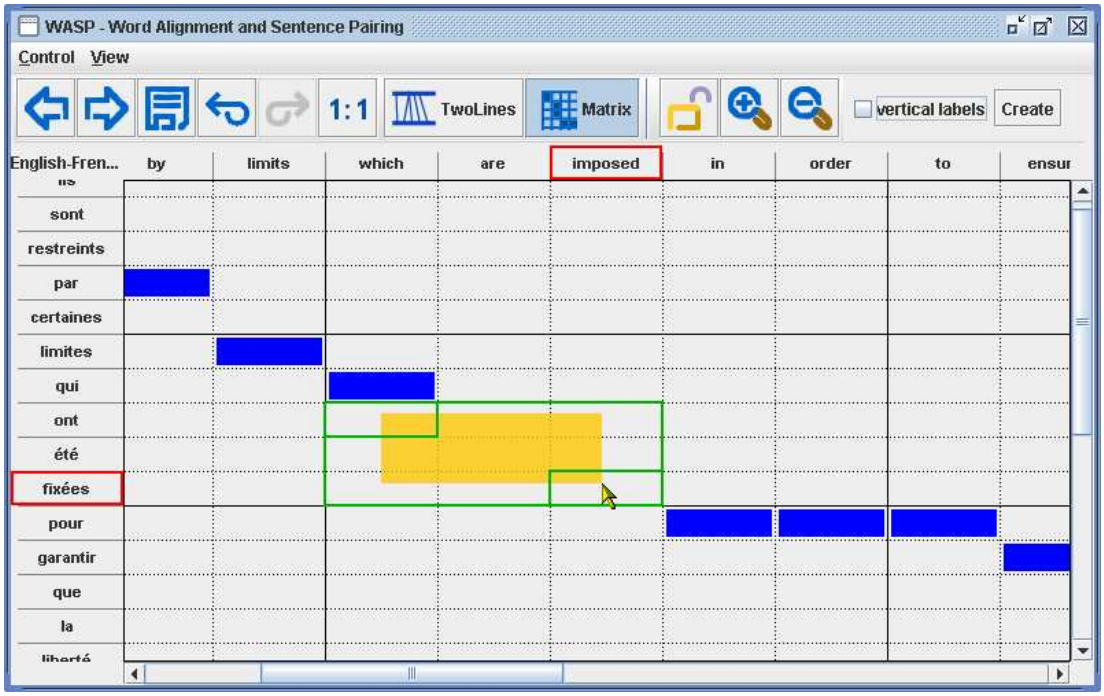
V pravém horním rohu aplikace (na panelu nástrojů pro pohled) je tlačítko zobrazující aktuální mód pro frázová spojení. Přepíná se kliknutím (nebo stiskem klávesy „M“) mezi módy *Create* a *Delete* viz. Obrázek 12. V příslušném stavu se pravým tlačítkem myši vybere souvislá oblast vyjadřující v obou větách příslušnou frázi.

Při uvolnění tlačítka myši dojde v módu *Create* k vyznačení fráze. V módu *Delete* se pak smažou všechna spojení ve vybrané oblasti (tedy ne nutně jen fráze).



Obrázek 12 - Výřez z pohledu Matrix (přepínání módu)

Nastavený mód je také zřetelný při samotném použití, kdy plocha, která bude obsazena párováními resp. smazána je ohraničena zeleným resp. červeným rámečkem. Ukázka vytváření frázového spojení - Obrázek 13



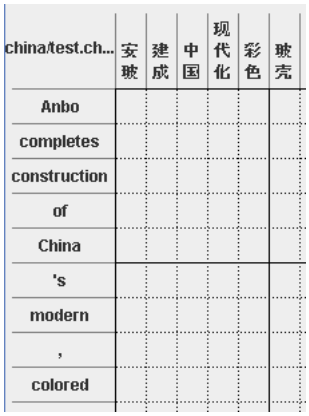
Obrázek 13 - Vytváření frázového spojení (maticové zobrazení)

Volitelné natočení slov v záhlaví sloupců umožňuje vytvořit tabulku se čtvercovými políčky, čímž se zvětšuje užitná plocha pro párování. Ve vodorovném směru se tak vejde na obrazovku větší část věty. Drobnou nevýhodou je snížená čitelnost takového zápisu, pokud je psán běžnou latinkou. Zajímavostí je, jak se toto natočení chová k exotickým jazykům jako je např. čínština. Komponenta, která zajišťuje vykreslování slov v záhlaví tabulky má v sobě zabudovánu kontrolu použitých znaků a v případě, že objeví v řetězci exotické znaky, zabrání jejich

nesprávnému natočení (ve výchozím provedení o 90° proti směru hodinových ručiček) a vybere např. vertikální řazení znaků, kdy znaky jsou psané jednotlivě pod sebe, bez natáčení. Tím je zajištěna správná čitelnost jak lze vidět na obrázcích 14-15. V plovoucím popisku se však tato slova vypisují vždy vodorovně.

china/test.ch...	安玻	建成	中国	现代化
Anbo				
completes				
construction				
of				
China				
's				
modern				
,				
colored				

Obrázek 14 - zobrazení čínských znaků (bez natočení)



china/test.ch...	安玻	建成	中国	现代化	彩色	玻壳
Anbo						
completes						
construction						
of						
China						
's						
modern						
,						
colored						

Obrázek 15 - zobrazení čínských znaků (s natočením)

Funkce lupy u tohoto zobrazení mění velikost písma použitého v záhlaví a upravuje šířku buněk dle aktuálně nastavené úrovni zvětšení / zmenšení.

5 Možnosti dalšího vývoje aplikace

Zde přináším několik dalších možností, jak by aplikace mohla být v budoucnu rozšířena. V rámci stávajícího vývoje se na tyto prvky nedostalo, neboť náročnost jejich realizace by neodpovídala očekávanému přínosu nebo myšlenka jejich zapracování přišla příliš pozdě. Jde o tyto náměty:

- Rozšíření možnosti na natáčení textu v Matrix view na cca 30-60°, aby vznikl rozumný kompromis mezi čitelností po řádcích a rozšířením pokrývané plochy díky zúžení jednotlivých sloupců. Bude potřeba vyřešit oddělení takto natočených slov (aby bylo dobře viditelné, jak jsou napojena na vykreslovanou matici).

- Rozšíření podpory ovládání maticového zobrazení s pomocí klávesnice a snížení závislosti na využití myši, které je v řadě případů pomalejší. Použitá komponenta vlastního návrhu neobsahuje zabudovanou podporu pro výběr spoje a jeho vytváření nebo rušení a závisí tak plně na využití myši.

- Podpora pro různě spolehlivá spojení (sure, possible), která je již částečně připravena pomocí vnitřního atributu „confidence“ každého spojení (s nastavenou hodnotou sure). Ještě je třeba přidat podporu do samotného uživatelského rozhraní (vytváření a vizualizace) a začlenit do jádra aplikace do metod pro ukládání a načítání párování.

6 Příručka pro uživatele

6.1 Požadavky na systém

Java Runtime spolu s aplikací vyžaduje alespoň 32MB volné paměti.

3 MB volného místa na disku pro aplikaci

Spuštěné grafické rozhraní operačního systému.

Nainstalovaná Java 2 SE Runtime ve verzi 1.5 a vyšší (k získání na adrese http://java.sun.com). Výchozí cesta ke spustitelným souborům Javy nastavená na systému.

myš, nebo jiné polohovací zařízení. Testováno bylo pouze s použitím myši.

Program byl úspěšně spuštěn a provozován na počítači s procesorem na taktu 733MHz a 256MB paměti RAM.

6.2 Instalace

Instalace programu se provádí zkopírováním obsahu adresáře s distribucí na požadované místo na cílovém stroji. Cílový adresář musí mít práva pro čtení případně i pro spouštění (při použití skriptů ke spuštění) pro všechny uživatele, kteří budou s programem pracovat. Je vhodné přidat adresář s programem také do výchozí cesty (v systému označené jako proměnná prostředí PATH), aby bylo možné program spouštět použitím zkráceného názvu „wasp“. Tomu pomáhají batchové a shell skripty ve složce s programem určené pro systémy Windows a systémy na bázi UNIXu. Na Unixových systémech lze navíc vytvořit alias například pro slovo `wasp`, které bude asociované s příkazem.

```
java –jar <cesta k programu>/WASP.jar
```

Tímto voláním se následně program také spouští.

6.3 Nastavení programu

Program se snaží načíst maximum nastavení z konfiguračního souboru, který hledá nejprve v aktuálním adresáři. Konkrétně v souboru *„settings.ini“*. Tam je možné určit zdroj vět (soubory) pro jednotlivé jazyky, kódování těchto souborů, oddělovače slov a také cílový adresář, kam se budou ukládat výsledná párování. Výchozí umístění je aktuální adresář, ze kterého je aplikace spuštěna. Nastavení je

tedy umístěno u zdroje vět a také u výsledku párování. Pomocí dodatečného parametru při spouštění programu lze změnit očekávané umístění souboru s nastavením. Cesty v souboru s nastavením pak musí být buď absolutní, nebo vztažené k adresáři s nastavením.

Při poskytnutí veškerých důležitých parametrů v příkazové řádce lze spustit program i bez využití konfiguračního souboru. Tento postup se však nedoporučuje.

6.4 Struktura souboru nastavení

Soubor s nastavením je jednoduchý textový soubor, ve kterém jsou uloženy dvojice vlastnost a její hodnota. Pro zadávání stavů zapnuto / vypnuto u vlastností je využito binární stupnice, kdy 1 značí zapnutí funkce a 0 její vypnutí. Řádky, které nemají být použity je třeba převést na komentář použitím znaku „#“ na začátku řádku.

<název-vlastnosti>=<**výchozí_hodnota** | hodnota2 | ...>

antialiasing = 1 | 0
Ovlivňuje použití jemnějšího vykreslování. Všechny čáry symbolizující párování jsou pak hladší a nemají tak ostré hrany. Cenou za to může být horší grafický výkon.

context-size = -1 | 0 … 100
Nastavuje délku zobrazovaného kontextu vpřed a nazpět v aktuálním bloku vět. Maximální délka je omezena na 100 vět, neboť delší kontext se pravděpodobně vyskytovat nebude a pokud by někdo chtěl využít delší, může nastavit neomezený kontext.

- 0 odpovídá vypnutému zobrazování kontextu.
- -1 neomezenému kontextu, kdy se zobrazuje, vše co lze.

lang1-sentpair = <soubor-se-seznamem>
Nastavuje soubor, který v sobě obsahuje seznam jednotlivých bloků k párování. Na každém řádku musí být uvedena cesta k jednomu souboru s blokem vět pro první jazyk. Cesta k souboru musí být relativní vůči souboru s nastavením nebo absolutní.

Alternativně je možné zadat seznam přímo do tohoto nastavení ve formátu

```
:soubor_1:soubor_2: ... :soubor_N:
```

to se může hodit pro malý počet souborů, kdy uživatel nechce pro seznam vytvářet nový soubor.

lang1-encoding =

Určuje kódování vstupního souboru s větami. Při neuvedení se jako výchozí nastavení použije výchozí kódování Javy, které odpovídá výchozímu systémovému kódování na dané platformě.

lang1-separator = [ ]

Nastavuje oddělovač slov v daném jazyce. Umožňuje nastavit použití zvláštního tagu jako oddělovače. Hodnota musí být zadána s ohledem na vlastnosti regulárních výrazů v Javě. Pozor při zadávání bílých znaků, které se nemusí správně uložit a interpretovat při načítání. Je třeba je uzavřít do patřičného druhu závorek nebo zadat jako escape sekvenci „_“

Nastavení pro druhý jazyk se provede obdobně, jen u názvu vlastnosti se použije na začátku „lang2“ místo „lang1“

null-connections = 1 | 0

Určuje, zda lze v aplikaci vytvářet prázdná spojení pro slova bez ekvivalentu v druhém jazyce.

tagged = 0 | 1

Určuje, zda jsou věty uvozeny identifikačními tagy. Tedy ve tvaru:

```
<s id="identifier">The sentence
```

nebo i jiném podobném formátu. Těsně souvisí s vlastností *sentence-tag-regexp*.

sentence-tag-regexp =

Pokud je věta značkovaná, určuje tento parametr formát použitých značek. Musí respektovat pravidla pro regulární výrazy v Javě. Lze využít některou z připravených sad. Pro typické formáty značek. Pro značky uvedené výše to může být:

```
<s [^>]*id="([^"]+)"[^>]*>(.*)
```

pro věty ve tvaru: `<s id=“identifikator“>obsah`

sentence-tag-id-group =

Určuje číslo skupiny, dle pravidel *capture groups* regulárních výrazů, která obsahuje identifikační řetězec věty z vlastnosti *sentence-tag-regexp*

sentence-tag-content-group =

Označuje číslo skupiny ve vlastnosti *sentence-tag-regexp*, která obsahuje samotné čisté znění věty.

output-separator = <space>

Výstupní oddělovač sloupců v souboru s párováními. Odděluje jednotlivé indexy slov. Výchozí nastavení je mezera.

output-encoding = utf-8 | cp1250 | iso8859-2 | …

Kódování souborů s uloženými párováními. Projevuje se hlavně při souběžném ukládání párovaných slov.

output-dir = <aktuální adresář>

Označuje výstupní adresář, kam se budou ukládat soubory s párováními a statistiky. Výchozí je aktuální adresář resp. adresář s uloženým nastavením.

word-output-template =

Šablona, do které budou při ukládání párování dosazena znění slov, která pár spojuje. Ve výchozím nastavení se slova s párováním neukládají. Slova budou na pozicích označených *%s*. Doporučená hodnota šablony je: (%s, %s)

force-creation = 1 | 0

zda kolize obyčejného spojení s *Null connection* bude hlášena jako chyba nebo bude přepsáno na novou hodnotu. Výchozí nastavení je odstranit kolizi a provést požadovanou operaci.

maximum-sentence-time = 0

Umožňuje nastavit maximální čas v sekundách strávený na větě. Po jeho vypršení je uživatel informován o uplynutí limitu a je vyzván k uložení práce a přechodu na další větu. Překročení času je zaznamenáno do statistik. Nastavení limitu na 0 vypíná tohoto omezení. Maximální hodnota tohoto parametru je omezena spíše použitím. Jedná se o statické omezení, které nebere v úvahu délku věty a tedy i složitost věty.

show-elapsed-time = 0 | 1

zda se bude na hlavním formuláři zobrazovat uplynulý čas párování věty.

maximum-button-width = 0 | ...

nastavuje maximální délku slov ve vizualizacích.Hodnota se zadává v pixelech a je třeba trochu experimentovat s nastavenou hodnotou, aby vyhovovala vlastnostem párovaných textů (průměrná délka slov, ...)

Nastavení limitu na 0 odpovídá výchozímu nastavení zrušení limitu (u maticového zobrazení se nadále limit uplatňuje a je nastaven na přibližně 10 znaků)

log-usage = 0 | 1 | 2

Určuje, zda se budou vytvářet statistiky z běhu programu a jak budou podrobné. Čím vyšší číslo, tím vyšší podrobnost.

0 – odpovídá vypnutému výstupu statistik

1 – zaznamenává se

- počet vět, které uživatel modifikoval
- počet odchodů od aplikace (přerušení práce na 20s)
- počet vypršení časového limitu větu (viz. *maximum-sentence-time*)
- čas strávený u aplikace v minutách (do tohoto času se nepočítá doba, kdy uživatel nereaguje. Tedy kdy aplikace nedostává zprávy o vstupu uživatele pohyb kurzoru / vstup z klávesnice.

2 – vše předchozí a navíc: statistika využívání funkcí programu (klávesové zkratky, přechod po větách tlačítkem nebo pres „*Goto sentence*“ dialog, vytváření *Null connection*). Slouží k testování a pro optimalizaci programu dle využívání funkcí..

6.5 Ukázková data

Vedle programu samotného jsou připraveny také ukázkové sady dvojjazyčných textů párované na úrovni vět. Slouží k vyzkoušení aplikace a také ukazují použití konfiguračních souborů. Jedná se o ukázková data získaná z prací [3, 6-7]. U některých příkladů je k dispozici také jejich párování.

7 Použité pojmy

Null connection, Nullword, Nullpair

Zvláštní typ spojení pro slovo ve větě v jednom z jazyků ukazující, že slovo nemá ekvivalent v druhé větě.

Connection, Pair

Běžné spojení mezi slovy dvou vět různých jazyků. Symbolizuje vztah, že jedno slovo je překladem druhého.

Phrasal connection

Zvláštní typ běžného spojení, které je použito mezi vícero slovy jedné a případně i druhé věty různých jazyků. Symbolizuje ekvivalenci celých frází a takové spojení je symbolizováno úplným bipartitním spojením mezi všemi zapojenými slovy.

Sentence

Jedna věta ze vstupního souboru. Skládá se ze seznamu slov oddělených separátorem.

Sentpair

Ucelený blok vět jednoho jazyka. Lze předpokládat možnost kontextu mezi větami v rámci bloku.

Tag

Značka. Nejčastěji identifikátor uzavřený mezi špičaté závorky (<s> <sp> <doc>) případně obsahující metadata ve formě atributů.

8 Závěr

Tato bakalářská práce měla za úkol návrh a implementaci uživatelského rozhraní pro ruční párování slov dvojjazyčných textů (word alignment). Na základě objevených nedostatků obdobných aplikací pro tento účel je výsledkem práce grafický nástroj, který nemá nedostatky stávajících aplikací a přináší i některá další vylepšení.

Vyvinuté rozhraní s přepínacími pohledy neomezuje uživatele a umožňuje mu využít takové prostředí, ve kterém se mu úloha bude lépe řešit. Maticové zobrazení zjednodušuje vytváření i rušení spojení (nyní na jedno kliknutí). Řádkové zobrazení pak zjednodušuje mazání spojení, neboť stačí na spojnici kdekoliv kliknout a umožňuje také modifikovat již hotová spojení. Oba navržené pohledy nabízejí jednoduchý způsob pohybu nad větou s funkcí „ruky“, a odstraňují problém společného posunu.

Aplikace zjednodušuje také párování více bloků textu. Nastavení aplikace je uloženo v souboru a tím se zabrání možným chybám se záměnou pořadí souborů, kdy výstupní soubory s párováními ztrácí vazbu na správné zdrojové věty.

Generování zvláštních výstupů s dobou odvedené práce, počtem zpracovaných vět a počtem vytvořených spojení se v žádné ze studovaných aplikací nevyskytovalo. Přitom poskytuje doplňkové informace k výslednému párování a době za kterou ho bylo dosaženo. Umožňuje zefektivnit řešení úlohy ručního word alignmentu v situaci, kdy je na tento úkol najata externí pracovní síla. Umožňuje srovnávat jednotlivé anotátory a vybrat pro spolupráci takové, kteří jsou v párování výkonní.

9 Reference

[1] Ion, R., and Tufis, D. (2004). Multilingual Word Sense Disambiguation Using Aligned Wordnets. Romanian Journal on Information Science and Technology, Dan Tufiş (ed.), Special Issue on BalkaNet, Romanian Academy, 7 (1-2), 198-214.
http://www.ceid.upatras.gr/Balkanet/journal/17_SemanticValidation.pdf

[2] The EGYPT toolkit (2000), Developed by the Statistical Machine Translation team, WS'99, CLSP/JHU.
http://www.clsp.jhu.edu/ws99/projects/mt

[3] Rebecca Hwa & Nitin Madnani (2004), UMIACS Word Alignment Interface
http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm

[4] Ted Pedersen & Brian Rassier (2003), Aligner for Parallel Corpora
http://www.d.umn.edu/~tpederse/parallel.html

[5] P. Lambert and coll. (2005). Guidelines for Word Alignment Evaluation and Manual Alignment. Language Resources and Evaluation, 39 (4). pp. 267-285. Springer.
http://lilaproject.org/veu/LR/epps_ensp_alignref.php3

[6] I. Dan Melamed. (1998) "Manual Annotation of Translational Equivalence: The Blinker Project," Institute for Research in Cognitive Science Technical Report #98-07. University of Pennsylvania, Philadelphia, PA.

[7] Franz Josef Och, Hermann Ney. (2000) "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China

10 Přílohy

Obsah přiloženého CD

dist/	adresář s výslednou distribucí programu
dist/javadoc	vývojová dokumentace k programu
dist/WASP.jar	samotná aplikace
test_data/	adresář se vzorovými daty
src/	adresář se zdrojovými soubory
Bc_prace_Sochna_Jan.pdf	bakalářská práce