Paralelní korpusy

0/2 Z, zimní semestr 2006/2007

Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Filozofická fakulta Univerzity Karlovy v Praze

1 Různé

Filmové titulky I

http://www.opensubtitles.org/

http://divxsubtitles.net/

Filmové titulky II

1 / 00:01:15,708 –> 00:01:18,270
My name Borat. I like you.

2 / 00:01:19,037 –> 00:01:20,026
I like sex.

3 / 00:01:21,091 –> 00:01:22,309
It nice.

4 / 00:01:23,403 –> 00:01:25,399
This my country of Kazakhstan.

5 / 00:01:26,205 –> 00:01:31,818
It locate between Tajikistan and Kirghistan,
and assholes, Uzbekistan.

1 / 00:01:14,268 –> 00:01:18,949
Moje meno je Borat. Mám vás rád.

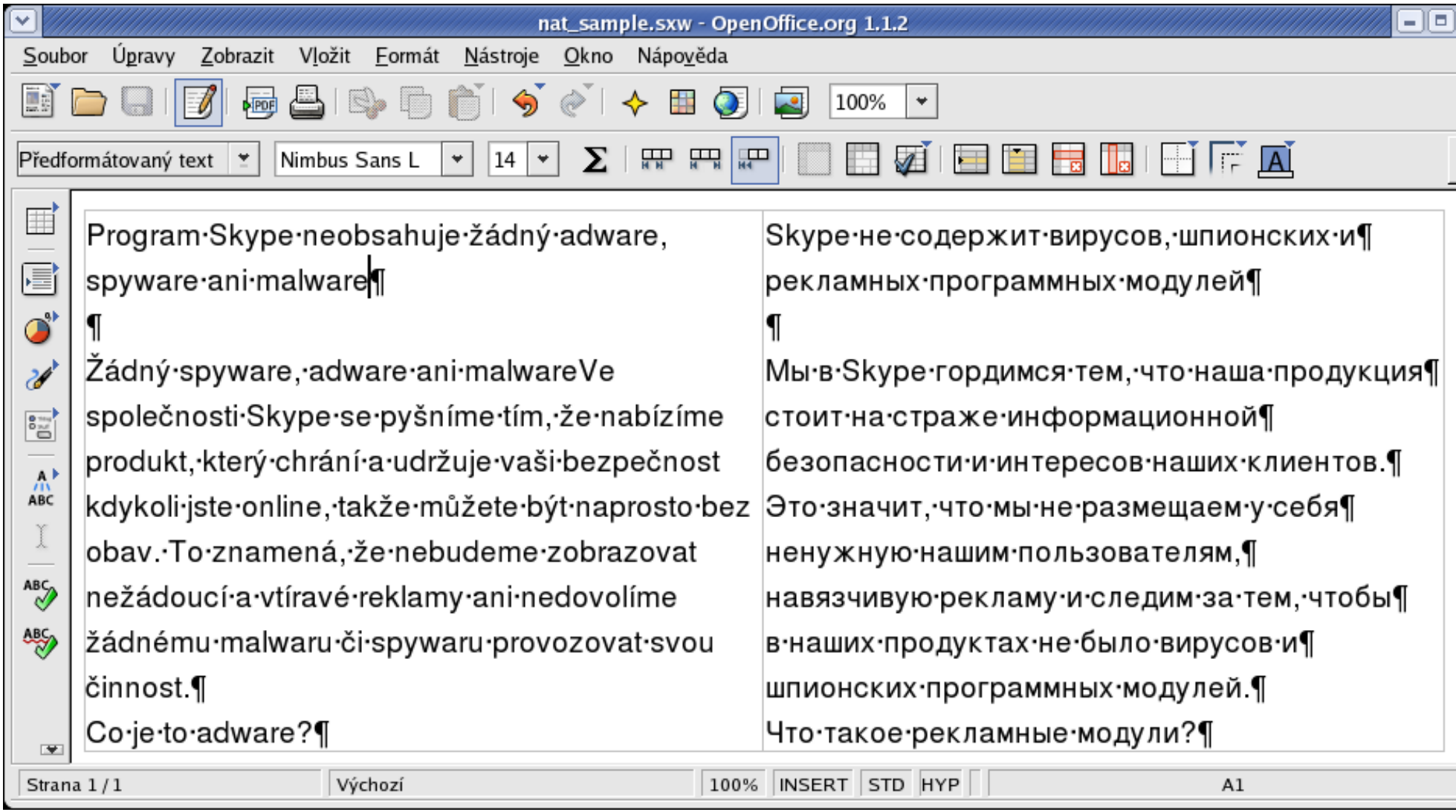
2 / 00:01:19,084 –> 00:01:19,919
Mám rád sex.

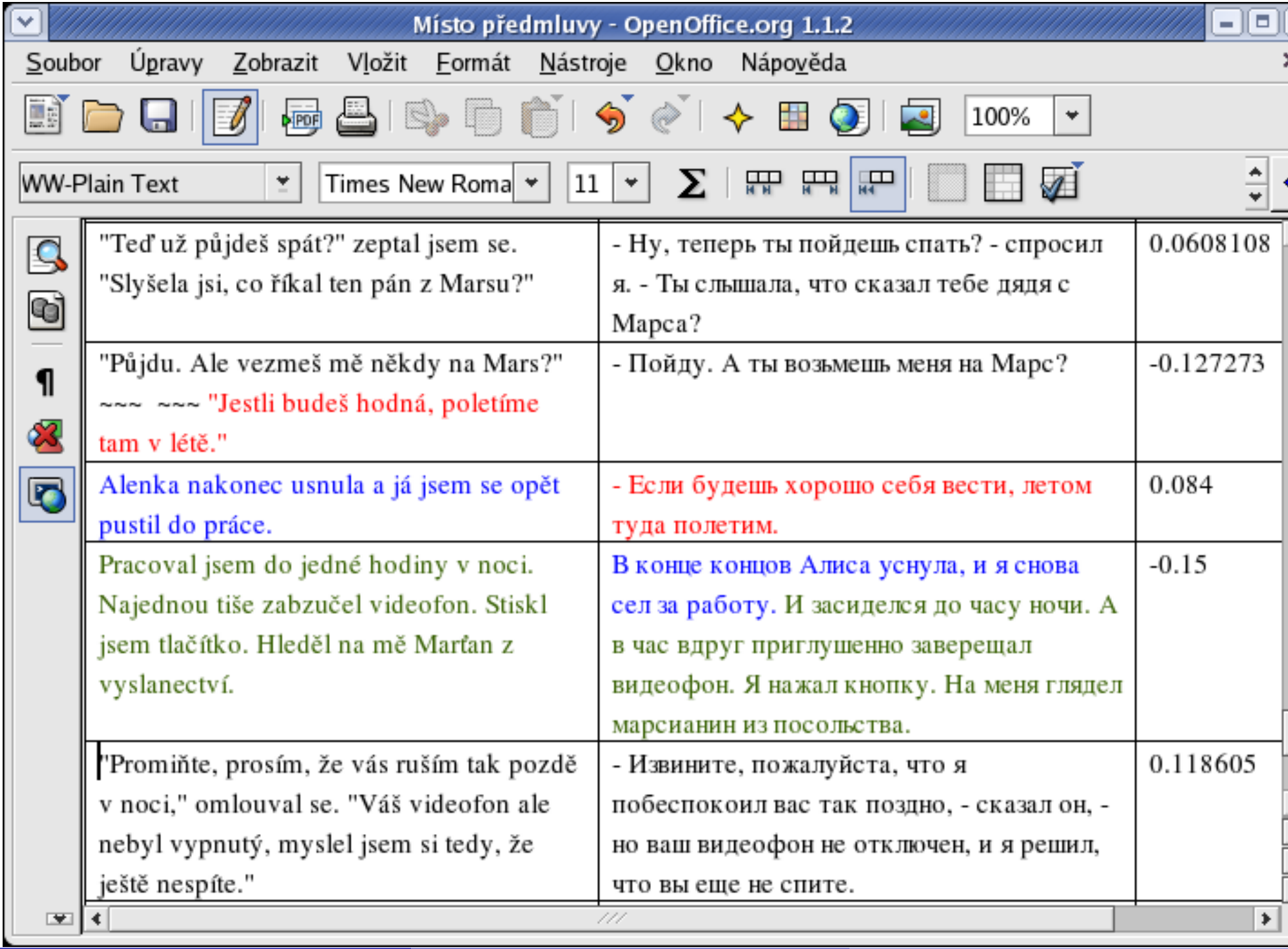
3 / 00:01:21,099 –> 00:01:22,299
Je hezký.

4 / 00:01:23,219 –> 00:01:25,819
Tohle je moje země, Kazachstán.

5 / 00:01:26,819 –> 00:01:31,819
Leží mezi Tádžikistánem, Kirgistánem
a prdelí světa - Uzbekistánem.

Problémy s formátem vstupu





"Teď už půjdeš spát?" zeptal jsem se. "Slyšela jsi, co říkal ten pán z Marsu?"	- Ну, теперь ты пойдешь спать? - спросил я. - Ты слышала, что сказал тебе дядя с Марса?	0.0608108
"Půjdu. Ale vezmeš mě někdy na Mars?" ~~~ ~~~ "Jestli budeš hodná, poletíme tam v létě."	- Пойду. А ты возьмешь меня на Марс?	-0.127273
Alenka nakonec usnula a já jsem se opět pustil do práce.	- Если будешь хорошо себя вести, летом туда полетим.	0.084
Pracoval jsem do jedné hodiny v noci. Najednou tiše zabzučel videofon. Stiskl jsem tlačítko. Hleděl na mě Marťan z vyslanectví.	В конце концов Алиса уснула, и я снова сел за работу. И засиделся до часу ночи. А в час вдруг приглушенно заверещал видеофон. Я нажал кнопку. На меня глядел марсианин из посольства.	-0.15
"Promiňte, prosím, že vás ruším tak pozdě v noci," omlouval se. "Váš videofon ale nebyl vypnutý, myslel jsem si tedy, že ještě nespíte."	- Извините, пожалуйста, что я побеспокоил вас так поздно, - сказал он, - но ваш видеофон не отключен, и я решил, что вы еще не спите.	0.118605

kundecsar.doc - Microsoft Word

""Prosím tě.	"إنه ألم ممض .
	يتوقعون منا أن نقيم حفلات الموسيقى مجانا ، بدون مقابل .
Pořád nás nutí, abychom vystupovali zadarmo.	وكل يوم يأتون بذريعة جديدة .
Jednou ve prospěch boje proti imperialismu, podruhé k výročí revoluce, potřetí k narozeninám nějakého potentáta, a když nechci, aby nás zlikvidovali, musím se vším souhlasit.	مرة من أجل الكفاح ضد الامبريالية ، وأخرى للعيد السنوي للثورة ، والمرة التالية نحتفل بميلاد ذاك العظيم ، إذا أردت الاحتفاظ بالفرقة معا ، فلابد أن أساير كل شيء .
	ليس عندك أي فكرة بماذا ضغطوا على اليوم ".
Nevíš, jak jsem se dnes zas rozčílil.	سألته راضية "ما هو؟" .
""Copak?	"امرأة من المجلس المحلي حضرت العزف وبدأت تحاضرنا عن المفروض أن نعزفه والمرفوض ، وفي النهاية مالقتنا بحفل موسيقى بالمجان للجنة الشبابية .
" řekla bez zájmu.	وذلك لم.
"Navštívila nás při zkoušce nějaká referentka z národního výboru a začala nás poučovat, co smíme hrát a co nesmíme, a nakonec nás přinutila, abychom uspořádali zadarmo koncert pro Svaz mládeže, ale to nejhorší je, že zítra musím celý den strávit na jakési pitomé konferenci, kde nás budou poučovat o tom, jak má hudba pomáhat budovat socialismus.	يكن أسوأ ما حدث – فسوف أقضي طول يوم غد في مؤتمر غبي ، يثرثرون فيه عن دور الموسيقى في بناء الاشتراكية .
Zkažený den, úplně zkažený den!	يوم كامل ذهب إلى الجحيم !

Struktura textu stejná jako v originále?

Ne nutně. Jazyky se liší v užívání:

- interpunkce
- dělení na věty
- přímé a nepřímé řeči

Příklad

– Izvinite, požalujsta, čto ja pobespokoil vas tak pozdno, – skazal on, – no vaš videofon ne otključen, i ja rešil, čto vy ešče ne spite.

"Promiňte, prosím, že vás ruším tak pozdě v noci," omlouval se. "Váš videofon ale nebyl vypnutý, myslel jsem si tedy, že ještě nespíte."

Zarovnávání textů s odlišnou strukturou

Předpoklady při zarovnávání:

1. shodné nebo nepatrně odlišné pořadí vět v paralelních textech
2. minimum přidaných nebo vypuštěných pasáží
3. většina vět odpovídá 1:1, v jiných případech jsou čísla v *m*:*n* nízká

– vše kvůli efektivitě

Příliš často neodpovídá realitě!

Řešení?

- úprava textů před zarovnáním
- načtení textů do databáze, hledání korespondencí bez ohledu na pořadí

Zarovnávání slov, výrazů, větných členů

Předpoklad:

1. segmentace/tokenizace v paralelních textech (nezávisle)
2. zjišťování korespondencí (zarovnání)

Segmentace ale může záviset na druhém jazyku:

- *patentová přihláška*
- *demande de brevet*
- *Patenanmeldung*
- *domanda di brevetto*

Řešení?

Víceúrovňová segmentace!

Drží se překladatelé co nejvíce originálu?

Záleží na typu textu. V beletrii spíše ne.

Důvody:

- cílový jazyk nemá srovnatelný výraz nebo konstrukci
- překladatel dá ze stylistických důvodů přednost jinému výrazu nebo konstrukci, i když má k dispozici „doslovnější“ variantu
- překladatel se bojí, že udělá chybu, když použije identické výrazové prostředky

A když vypadá překlad podobně jako originál –

– tak může jít o neumělý, nepřirozený, doslovný překlad

Navíc překladatelé někdy chybují

– a některé chyby může odhalit jen velmi dobrý znalec obou jazyků

Co když nemáme paralelní, ale jen srovnatelné texty?

Texty mohou být „neparalelní“ v různé míře:

- stejné věty jsou v textech na jiných pozicích
- texty obsahují jen větší či menší podíl stejných vět
- texty nepojednávají o stejném tématu
- texty nejsou ze stejného oboru

Výsledkem je, že:

- výrazu nelze vždy přiřadit jednoznačný překlad
- ne vždy lze z textů překlad zjistit
- četnosti ekvivalentních výrazů v textech nelze srovnávat

Jak využít srovnatelné texty I

Ale:

- je-li téma stejné, ekvivalentní výrazy se vyskytují ve všech jazycích ve srovnatelném kontextu
- v daném oboru a v určité době se ekvivalentní výrazy vyskytují se srovnatelnou frekvencí

Jsou-li texty ze stejného oboru, na stejné téma a ze stejné doby:

- ekvivalentní výrazy se vyskytují v podobných kontextech
- ekvivalentní výrazy jsou srovnatelně frekventované

Jak využít srovnatelné texty II

Hledání ekvivalentu podle srovnatelného kontextu

1. vyhledat slovo S_A s kontextem v jazyce A
2. přeložit slova v kontextu S_A pomocí *nějakého* slovníku do jazyka B
3. vyhledat kontexty s přeloženými slovy v jazyce B
4. hledané slovo S_B je to, které je v těchto kontextech nejčastější

Jinak řečeno... (parafráze) I

K čemu jsou parafráze dobré:

- generování (syntéza) přirozeného jazyka
- sumarizace
- hodnocení systémů strojového překladu
- hodnocení dotazovacích systémů

Využití jednojazykového paralelního korpusu k parafrázování

Emma **burst into tears** and he tried to **comfort** her, **saying things to make her smile**.

Emma **cried**, and he tried to **console** her, **adorning his words with puns**.

Jinak řečeno... (parafráze) II

Postup

1. zarovnání po frázích (skupinách slov)
2. *This situation is . . . in terms of security*
3. *under control* $\longrightarrow$ *unter Kontrolle*
4. *unter Kontrolle* $\longrightarrow$
   - *in check*
   - *checked*
   - *curbed*
   - **curb*
   - **limit*
   - **slow down*

(Bannard & Callison-Burch, ACL 2005)