Description of the *Le Monde Diplomatique-Die Tageszeitung* Parallel Corpus

Garance Paris

Version 0.1
23rd April, 2008

Acknowledgements: I wish to thank Elke Teich and Silvia Hansen for their help while I was building this corpus, Erich Steiner and Andrea Kamm for providing the funding for the aligning the texts, and Véronique Nessius for her help with the alignement.

The texts composing the *LMD-TAZ* French-German parallel corpus were taken from the CD-Rom archive of the French monthly newspaper *Le Monde Diplomatique* (2001), of which a German edition is also available, published by *taz, die tageszeitung*.[1] Only those articles which were accompanied by a German translation on the CD-Rom were selected.

In its present state, the corpus comprises the relevant articles from the January, February, March, October, November and December 1999 issues, altogether 136 articles in each language. It contains almost 243 000 word tokens for French and a little over 224 000 for German. For information, some corpus statistics are displayed below:

	French	German
Tokens	242 695	224 230
Types	20 700	27 747
Sentences	9 385	9 218

The alignment units are in general full sentences, i. e. main clauses containing a finite verb together with their dependent subordinate clauses. Material which was missing in one of both versions of the text was deleted, and in some cases where the splitting or joining of sentences in the two versions of the texts made aligning very difficult, the material was removed. During the alignment process, small corrections were made to the text, such as rectifying typos, and the files were later spell-checked to improve tagging accuracy.

[1] See `http://www.taz.de`.

The files have been tokenised and tagged with the TreeTagger[2] (Schmid, 1994, 1995; Stein and Schmid, 1995) and its parameter files for French and German, which use the Stuttgart-Tübingen (Schiller et al., 1995) and Stein (Stein, 1995) tag sets. The lemma of each word is printed next to it and its assigned PoS tag, as shown in Figure 1.

A Parallel Corpus, but not a Translation Corpus

It should be mentioned that the resulting corpus is a parallel corpus, but that it cannot really be considered as a translation corpus, because we cannot properly speak of "originals" and "translations" (or of "source texts" and "target texts") when referring respectively to the *Le Monde Diplomatique* and the *TAZ* versions of the texts, the articles being more similar to parallel versions of the same stories than to translations from one language into another.

Indeed, there are often whole sentences, sometimes even whole paragraphs, which are present in only one of both languages, be it French or German, as can be seen in the excerpt shown in Figure 2. Moreover, there are also sometimes differences in the figures, dates, etc. in both versions of an article (see Figure 3). This is due due a thorough rewriting of the German text after it has been translated from French, resulting in two parallel versions of the same text.

References

Le Monde Diplomatique (2001). Archives 1980-2000. 21 années du Monde Diplomatique en texte intégral. CD-Rom. Order from `http://www.monde-diplomatique.fr/cederom/`.

Schiller, A., Teufel, S., and Stöckert, C. (1995). Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/stts_guide.ps.gz`.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK. Revised version of the paper available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.ps.gz`.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop*. Revised version of the paper available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.ps.gz`.

Stein, A. (1995). Liste der Tags in französischen Korpora. Available from `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`.

Stein, A. and Schmid, H. (1995). Étiquetage morphologique de textes français avec un arbre de décisions (Morphological tagging of French texts using a decision tree). *t.a.l. (Traitement Automatique des Langues)*, 36 (Traitement probabilistes et corpus—Probabilistic Approaches to Corpora)(1–2):23–36. Available from `http://www.uni-stuttgart.de/lingrom/stein/pubs/steins95.ps.gz`.

[2]The TreeTagger and its parameter files for German and French are freely available for download from `http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html`

<seg_id=1>			
Quelles	PRO:REL	quel	fr
sont	VER:pres	être	fr
,	PUN	,	fr
à	PRP	à	fr
l’	DET:ART	le	fr
aube	NOM	aube	fr
d’	PRP	de	fr
un	DET:ART	un	fr
siècle	NOM	siècle	fr
nouveau	ADJ	nouveau	fr
,	PUN	,	fr
les	DET:ART	le	fr
principales	ADJ	principal	fr
caractéristiques	NOM	caractéristique	fr
de	PRP	de	fr
la	DET:ART	le	fr
planète	NOM	planète	fr
?	SENT	?	fr
</seg>			
<seg_id=1>			
An	APPR	an	de
der	ART	d	de
Schwelle	NN	Schwelle	de
zum	APPRART	zum	de
neuen	ADJA	neu	de
Jahrhundert	NN	Jahrhundert	de
scheint	VVFIN	scheinen	de
es	PPER	es	de
angebracht	ADJD	angebracht	de
,	$,	,	de
sich	PRF	er|es|sie|Sie	de
Gedanken	NN	Gedanke	de
über	APPR	über	de
die	ART	d	de
Hauptmerkmale	NN	Hauptmerkmal	de
der	ART	d	de
heutigen	ADJA	heutig	de
Welt	NN	Welt	de
zu	PTKZU	zu	de
machen	VVINF	machen	de
.	$.	.	de
</seg>			

Figure 1: Example sentence. The first column is the word in French or German, the second its part-of-speech tag, the third its lemma, the fourth the language of the text.

La loi sur la sécurité nationale, dans son article 7, prévoit des peines de prison ferme pour toute activité "anti-étatique"*, ce qui inclut les prises de position favorables au régime de Pyongyang ou à l'idéologie communiste - "une violation flagrante des libertés fondamentales", souligne M. Kwak Nohyun.*	Das Gesetz zur nationalen Sicherheit sieht in Artikel 7 für jede "gegen den Staat" gerichtete Handlung Gefängnisstrafen ohne Bewährung vor.
⋮	⋮
	Nach einer weiteren Phase höchster Spannungen zwischen den beiden Ländern hat Washington Mitte September bekanntgegeben, dass das seit sechsundvierzig Jahren andauernde Wirtschaftsembargo gegen Nordkorea teilweise aufgehoben werden soll. Zuvor hatte sich das Regime in Piöngjang bereit erklärt, seine Tests mit Langstreckenraketen auszusetzen. Von einer Entspannung ist allerdings im Lande selbst nicht viel zu spüren; dort kämpfen die Menschen nach wie vor ums nackte Überleben, während das Regime "Normalitt" simuliert und sich mit einem dicken Wall aus Propanganda und Misstrauen umgibt. Aber auch Südkorea tut sich schwer mit dem Ende des Kalten Krieges.
De Séoul, plusieurs agences de tourisme offrent la possibilité de jeter un coup d'oeil sur le pays ennemi - la Corée du Nord - en organisant des visites de la DMZ, la zone démilitarisée qui sépare les deux Corées.	In Seoul organisieren mehrere Reiseveranstalter Ausflüge in die entmilitarisierte Zone (DMZ), die Nord- und Südkorea voneinander trennt und von der aus man einen Blick auf das feindliche Nachbarland werfen kann.
⋮	⋮
15 août 1948. La proclamation de la République de Corée (au sud) met officiellement fin à l'occupation américaine. Syngman Rhee est élu président.	15. August 1948: Die Proklamation der Republik Korea (im Süden) beendet offiziell die US-amerikanische Besatzung*, obgleich 35 000 amerikanische Soldaten auf der Halbinsel verbleiben.* Syngman Rhee wird zum ersten Präsidenten gewählt.
9 septembre 1948. Proclamation de la République populaire démocratique de Corée (RPDC), dont Kim Il-sung devient président.	9. September 1948: Proklamation der Demokratischen Volksrepublik Korea (DVRK). Präsident wird Kim Il Sung*, der dieses Amt bis zu seinem Tod im Jahre 1994 ausübt.*

Figure 2: Missing segments in both the French and German versions of a November 1999 article series on North and South Korea (the missing parts are printed in italics)

L'arrivée récente d'indigènes awas sur la terre qu'ils ont démarquée a provoqué des conflits (*six* maisons brûlées notamment).	Als Awa-Indianer ein Stück Land für sich absteckten, gingen *sieben* Häuser in Flammen auf.
En 1992, le rapport du groupe de travail de l'AP insista en ce sens et, en juin 1995, un autre rapport du groupe de travail prit position en faveur de la création des unités de visites familiales.	In einem Bericht von *Juni 1995* sprach sich auch eine Arbeitsgruppe der Abteilung Strafvollzug in diesem Sinne aus; in einem weiteren Bericht setzte sie sich für die Schaffung von Langzeitbesuchsräumen ein.

Figure 3: Two excerpts from February 1999 articles, which contain different figures and dates (in italics) in the French and German versions