Le corpus « Droits de l'Homme » du LLI

Christine CHODKIEWICZ, Fabrice ISSAC et Bénédicte PINCEMIN

Nous avons entrepris la construction d'un corpus électronique de textes juridiques des Droits de l'Homme. Celui-ci est actuellement constitué de 28 Conventions internationales en deux ou trois langues 'et comprend quelques 250 000 mots (version 1.0, mars 2005).

Il peut être téléchargé à l'adresse suivante :
`http://www-lli.univ-paris13.fr/ressources`

1. Configuration informatique du corpus

Destiné à devenir une ressource de référence en matière d'information juridique électronique, ce corpus se veut exemplaire quant à sa composition, sa structuration et son codage. Le choix des textes a été fait en lien avec Jean-Bernard MARIE[1]. Le standard international pour la diffusion de corpus (TEI) a été appliqué non sans avoir consulté la communauté scientifique sur l'expérience acquise en matière de codage de corpus juridiques (questions soumises sur la liste électronique CORPORA, contacts pris avec des éditeurs spécialisés). Les fichiers ont été collectés sur internet en veillant à la qualité des sources (cf. infra) puis normalisés au format XML et enfin édités sous la forme d'un corpus structuré au format TEI.

Il n'a jamais été développé de codage spécialement adapté à la nature particulière de ces textes. Aussi avons nous défini une DTD (Document Type Definition) permettant de rendre compte au plus près de la structure des conventions (préambule, sections, articles, dispositions finales etc…) tout en étant automatiquement transposable dans les structures plus générales prévues par la TEI. Le corpus existe ainsi en deux versions : une version TEI pour la diffusion (la TEI représente une quasi norme et propose une large gamme de balises) et une version selon la DTD spécifique pour l'archivage qui assure un meilleur contrôle de la structure des textes (articulation de parties de différentes natures : Préambule, articles, dispositions finales…) Notre objectif à terme est de créer un ensemble de DTD suffisamment précise et souple pour pouvoir représenter des textes juridiques qui ne seraient pas limités aux conventions ni au domaine des droits de l'homme.

La réalisation de l'entête TEI a également donné lieu à une documentation soignée du balisage (partie `<tagUsage>`), d'une importance cruciale pour une juste interprétation des traitements

[1] Directeur de recherche au C.N.R.S., Ancien Secrétaire général de l'Institut international des droits de l'homme, Strasbourg

effectués sur le corpus et pour des réutilisations dans d'autres contextes. Ce point mérite d'être souligné car un sondage des pratiques actuelles montre que bien souvent cette partie de l'entête se réduit à un inventaire quantitatif des balises. L'enjeu scientifique est donc d'illustrer l'utilité d'une description précise du codage tout en en démontrant la faisabilité (les exemples rencontrés étant pour la moitié les quelques cas d'école très brefs donnés dans les manuels et les guides).

L'entête prévoit la possibilité de déclarer une taxonomie (thésaurus, ontologie) à laquelle rapporter des mots-clés. La structuration intertextuelle du corpus a conduit à l'introduction de plusieurs taxonomies de référence, permettant de caractériser chaque texte aux plans de sa source (instance émettrice : ONU, Conseil de l'Europe…), de son ancrage chronologique (date d'entrée en vigueur) et de ses thématiques dominantes.

En particulier pour la description thématique nous avons repris la classification de référence mise au point par Jean-Bernard Marie (cf. Infra). Complémentairement, la construction d'une « ontologie » des Droits de l'Homme est en projet. C'est une entreprise tout à fait novatrice. En effet, les documents sont aujourd'hui organisés par des plans de classement, obligeant à quadriller artificiellement le domaine. Le passage d'une logique de classement à une logique d'indexation conduit à repenser toute la description sémantique du domaine, d'une façon beaucoup plus respectueuse des thèmes, sans plus être contraint par la recherche d'un équilibre contingent des classes ni durcir des associations récurrentes mais non intrinsèquement fixées.

Le corpus est trilingue – français/anglais/espagnol – pour les Conventions ONU, et bilingue – français/anglais – pour les autres conventions. Les conventions sont alignées au niveau de subdivisions fines (articles ou items de liste) via l'usage d'identificateurs structurés : chaque position dans le corpus est repérée par un indice et les indices, d'une langue à l'autre, se correspondent. Nous n'avons pas recouru à l'alignement par pointeurs et liens, plus lourd à mettre en œuvre et non nécessaire ici dans le domaine juridique, le strict parallélisme des structures jusqu'aux plus fines étant de rigueur (à un article correspond un article ; à une phrase correspond une phrase…)

Dès ses versions préliminaires, le corpus a fait l'objet de traitements automatisés : textométrie via WEBLEX[2] , mais aussi consultation des versions alignées à l'aide d'un concordancier multilingue (français/anglais et, lorsque nous disposons de textes en espagnol, des versions français/anglais/espagnol).

[2] développé par Serge Heiden (ICAR, C.N.R.S., FRE 2690, Institut de Linguistique Française, E.N.S. Lyon)

L'étiquetage morphosyntaxique du corpus est prévu à l'aide d'un logiciel, notre choix se portera sans doute sur CORDIAL ou la version française du Tree Tagger, toujours en respectant la TEI.

2. Précisions juridiques sur le corpus

2.1. Repères

Selon une classification de référence mise au point par Jean-Bernard Marie[3] ', on distingue en la matière de Droits de l'Homme, les conventions suivantes :

- **les conventions générales** qui intéressent l'ensemble des droits de l'homme ou un large groupe de ceux-ci (elles ont une portée **universelle ou régionale**)
- **les conventions spécifiques** qui visent à garantir certains droits de l'homme en particulier et qui concernent : le génocide, les crimes de guerre et crimes contre l'humanité ; l'esclavage, la traite des être humaines, le travail forcé ; la torture ; les disparitions ; l'asile ; la nationalité ; la liberté de l'information ; la vie privé, la sécurité sociale ; la bioéthique
- **les conventions relatives à la protection catégorielle** qui correspondent à la nécessité de protéger spécialement certaines catégories d'êtres humains : les réfugiés ; les apatrides ; les migrants ; les minorités ; les peuples indigènes ; les travailleurs ; les femmes ; les enfants ; les personnes avec un handicap ; les combattants, prisonniers et personnes civiles en temps de conflit armé
- **les conventions relatives aux discriminations** qui ont pour objet la lutte contre : la discrimination fondée sur la race ; la discrimination fondée sur le sexe ; la discrimination dans l'enseignement ; la discrimination dans l'emploi et la profession

2.2. Composition

Figurent dans notre corpus les conventions suivantes :

- Les convention générales à portée universelle que sont la **Déclaration universelle des droits de l'homme de 1948** et les deux **Pactes des Nations Unies de 1966** (Pacte international relatif aux droit économiques, sociaux et culturels et Pacte international relatif aux droits civils et politiques).

- Les conventions générales à portée régionale que sont la **Convention européenne des droits de l'homme de 1950** et la **Charte sociale européenne de 1961**, s'agissant de

[3] Instruments internationaux relatifs aux droits de l'homme : classification et état des ratifications au 1er janvier 2003, RUDH 2003, p. 59 et s.

l'Europe et **La Convention américaine relative aux droits de l'homme (1969)**, **La Charte africaine des droits de l'homme et des peuples (1981)** et la **Charte arabe des droits de l'homme (1994)**, s'agissant d'autres régions du globe.

- Les conventions spécifiques ne sont pratiquement pas représentées dans notre corpus, à deux exceptions près sur la torture **:** l''une européenne (la **Convention européenne pour la prévention de la torture et des peines ou traitements inhumains ou dégradants**), l'autre inter-américaine (la **Convention interaméricaine pour la prévention et la répression de la torture**).

- Les Conventions relatives à la protection catégorielle sont, elles aussi, presque toutes absentes notre corpus, à l'exception de l'une provenant de l'Organisation Internationale du Travail (OIT) – la **Convention sur la liberté syndicale et la protection du droit syndical (qui protège les travailleurs)** - et d'une autre de l'ONU – la **Convention relative aux droits de l'enfant** (qui, comme son nom l'indique, protège les enfants).

- Les Conventions relatives aux discriminations ne sont, quant à elles, représentées dans notre corpus que par une seule convention de l'ONU : la **Convention sur l'élimination de toutes les formes de discrimination à l'égard des femmes**.

La liste complète des conventions du corpus figure en annexe du présent document.

2.3. Délimitation du corpus : un choix sélectif mais motivé

On pourra objecter que les instruments retenus pour constituer le corpus DH du LLI constitue une sélection fort restrictive qui ne rend pas compte du développement qu'ont connu les normes internationales des droits de l'homme à travers tous les instruments spécialisés et catégoriels élaborés depuis 50 ans.

De fait, le choix des conventions, réalisé en 2002 avec le concours de Jean-Bernard MARIE avait pour but initial de mettre à jour son *Glossaire des droits de l'homme* publié en 1981 (projet en cours). La composition du corpus est délibérément limitée, mais elle a été établie sur des critères « objectifs et logiques » (selon l'expression même de J.-B. MARIE) et répond, nous semble-t-il, aussi à des exigences non négligeables de praticabilité. Procéder à des choix plus extensifs aurait soulevé le problème d'une sélection par trop subjective et arbitraire, sujette à contestation ou suscitant pour le moins des réserves. Or, le corpus DH du LLI a été conçu d'emblée dans la perspective d'une large acceptation ; c'est à cette condition, nous

semble-t-il, qu'il peut faire œuvre utile et faciliter la compréhension en se fondant sur des bases exemptes de suspicion parce qu'établies de manière systématique.

Par ailleurs, l'on sait par expérience, que les instruments généraux des droits de l'homme fournissent véritablement l'essentiel du vocabulaire des droits de l'homme tel qu'il s'est développé au niveau international et que dans nombre de cas, les déclarations ou conventions spécifiques qui ont été élaborées parallèlement ou successivement ne contribuent pas de manière décisive à un enrichissement du langage des droits de l'homme. En général, les instruments spécialisés apportent des précisions sur les conditions concrètes et les modalités qui permettent la réalisation d'un droit particulier ou d'une catégorie de droits, plutôt qu'ils ne formulent de nouveaux concepts. Si l'on peut très certainement trouver certains exceptions, il demeure que la plupart des concepts, principes et termes fondamentaux sont exprimés dans les instruments généraux.

Le corpus DH du LLI est destiné à évoluer et à inclure ultérieurement divers instruments spécialisés sur les droits de l'homme (conventions, déclarations, recommandations particulières), voire de la jurisprudence ou de la doctrine… sans parler des textes rédigés au sein d'ONG telle qu'Amnesty International.

3. Les sources du corpus

Les textes des conventions qui figurent dans notre corpus proviennent de sites **institutionnels, les mieux à même d'assurer la fiabilité juridique et linguistique des textes**. En matière juridique, l'édition électronique des conventions ne fait pas foi, la présence de coquilles (jamais négligeables en droit) étant toujours possible par rapport à l'édition originale qui fait référence (toujours papier pour le moment)

Les principaux sites exploités sont :

- Pour les Conventions de l'ONU : http://www.un.org
- Pour les Convention de l'OIT : http://www.ilo.org
- Pour les Convention du Conseil de l'Europe : http://www.coe.int
- Pour les Conventions américaines et interaméricaines : http://cidh.oas.org
- Pour la Convention arabe : http://www.aidh.org
- Pour les statuts de la Cour Pénale Internationale : http://www.icc-cpi.int

4. Mise à disposition du corpus

Le corpus est disponible au format XML selon la DTD TEI à l'adresse suivante :
`http://www-lli.univ-paris13.fr/ressources`

La TEI est le standard international pour le codage de corpus et respecte la norme XML. Le corpus est ainsi très largement exploitable et diffusable. En effet techniquement de nombreux outils existent pour exploiter et transformer XML. Cela rend très accessible la conversion du corpus aux formats spécifique généralement requis par les applications TAL.
Le corpus DH a ainsi été adapté pour un concordancier multilingue[4] et pour l'outil WEBLEX[5].

Le corpus est téléchargeable sous forme d'une archive compressé, qui
inclut un certain nombre de fichiers propres au corpus lui-même :

- `convention.maitre.xml` : fichier principal faisant appel à tous les autres
- `convention.entity.dtd` : fichier de déclaration des entités de noms de fichier
- `convention.enteteteicorpus.xml` : entête global du corpus

Chaque convention utilise :

- `convention`*`i`*`.xml` : fichier principal de la convention i
- `convention`*`i`*`.entete.xml` : entête spécifique à la convention i
- `convention`*`i`*`.`*`lang`*`.xml` : texte de la convention i dans la langue lang

Les balises utilisées pour étiqueter les différentes conventions sont en fait un sous-ensemble de la TEI. Le fichier `tei-linaco.dtd`[6] contient la DTD utilisée pour valider le corpus.

Nous mettons en ligne la version 1.0 du corpus. Comme toute réalisation informatique cette version doit être corrigée, par exemple d'éventuels problèmes d'alignement, et améliorée, en ajoutant d'autres conventions. Toute remarque adressée à notre courriel (`corpus@lli.univ-paris13.fr)` sera donc la bienvenue.

5. Annexe : liste des conventions du corpus

x.c.hr.un.1948
Déclaration universelle des droits de l'homme

---

[4] Disponible à l'adresse `https://extranet-lli.univ-paris13.fr/DH/index.php`
[5] Outil en ligne de textometrie (statistique textuel) développé par Serge Heiden (http://weblex.ens-lsh.fr/wlx). Pour le moment l'accès à l'outil est protégé par un mot de passe.
[6] Nous avons ainsi nommé le fichier car il correspond à la TEI avec les modules additionnels sur les liens (li), les noms (na) et les corpus (co).

Universal Declaration of Human Rights

x.c.hr.al.1948a
Déclaration américaine des droits et devoirs de l'homme
American Declaration of the rights and duties of man

x.c.hr.al.1948b
Convention sur la liberté syndicale et la protection du droit syndical, 1948
Freedom of Association and Protection of the Right to Organise Convention, 1948

x.c.hr.ce.1950
Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales telle qu'amendée par le Protocole n° 11
Convention for the Protection of Human Rights and Fundamental Freedoms as amended by Protocol No. 11

x.c.hr.ce.1952
Protocole additionnel à la Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales, tel qu'amendé par le Protocole n° 11
Protocol to the Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocol No. 11

x.c.hr.ce.1961
Charte sociale européenne
European Social Charter

x.c.hr.ce.1963
Protocole n° 4 à la Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales, reconnaissant certains droits et libertés autres que ceux figurant déjà dans la Convention et dans le premier Protocole additionnel à la Convention, tel qu'amendé par le Protocole n° 11
Protocol No. 4 to the Convention for the Protection of Human Rights and Fundamental Freedoms, securing certain rights and freedoms other than those already included in the Convention and in the first Protocol thereto, as amended by Protocol No. 11

x.c.hr.un.1966a
Pacte international relatif aux droits économiques, sociaux et culturels
International Covenant on Economic, Social and Cultural Rights

x.c.hr.ce.1996a
Charte sociale européenne (révisée)
European Social Charter (revised)

x.c.hr.un.1966b
Pacte international relatif aux droits civils et politiques
International Covenant on Civil and Political Rights

x.c.hr.un.1966c
Protocole facultatif se rapportant au Pacte international relatif aux droits civils et politiques
Optional Protocol to the International Covenant on Civil and Political Rights

x.c.hr.al.1969
Convention américaine relative aux droits de l'homme
American Convention on Human rights, "Pact of San Jose, Costa Rica"

x.c.hr.un.1979
Convention sur l'élimination de toutes les formes de discrimination à l'égard des femmes
Convention on the Elimination of All Forms of Discrimination against Women

x.c.hr.ce.1983
Protocole n° 6 à la Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales concernant l'abolition de la peine de mort, tel qu'amendé par le Protocole n° 11
Protocol No. 6 to the Convention for the Protection of Human Rights and Fundamental Freedoms concerning the abolition of the death penalty, as amended by Protocol No. 11

x.c.hr.ce.1984
Protocole n° 7 à la Convention de sauvegarde des Droits de l'Hommeet des Libertés fondamentales, tel qu'amendé par le Protocole n° 11
Protocol No. 7 to the Convention for the Protection of Human Rights and

Fundamental Freedoms, as amended by Protocol No. 11

x.c.hr.al.1987
Convention inter-américaine pour la prévention et la répression de la torture Inter-American Convention to Prevent and Punish Torture

x.c.hr.ce.1987
Convention européenne pour la prévention de la torture et des peines ou traitements inhumains ou dégradants
European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment

x.c.hr.ce.1988
Protocole additionnel à la Charte sociale européenne
Additional Protocol to the European Social Charter

x.c.hr.al.1998
Statut de Rome de la Cour pénale international
Rome Statute of the International Criminal Court

x.c.hr.al.1988
Protocole additionnel à la Convention américaine relative aux droits de l'homme traitant des droits économiques, sociaux et culturels – « Protocole de San Salvador ».
Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social and Cultural Rights, "Protocol of San Salvador"

x.c.hr.un.1989a
Deuxième Protocole facultatif se rapportant au Pacte international relatif aux droits civils et politiques, visant à abolir la peine de mort
Second Optional Protocol to the International Covenant on Civil and Political Rights, aiming at the abolition of the death penalty

x.c.hr.un.1989b
Convention relative aux droits de l'enfant
Convention on the Rights of the Child

x.c.hr.al.1990

Protocole à la Convention américaine relative aux droits de l'homme traitant de l'abolition de la peine de mort

Protocol to the American Convention on Human Rights to Abolish the Death Penalty

x.c.hr.al.1994

Charte arabe des droits de l'homme

Arab Charter on Human Rights

x.c.hr.ce.1995

Protocole additionnel à la Charte sociale européenne prévoyant un système de réclamations collectives

Additional Protocol to the European Social Charter Providing for a System of Collective Complaints

x.c.hr.ce.1996b

Accord européen concernant les personnes participant aux procédures devant la Cour européenne des Droits de l'Homme

European Agreement relating to Persons Participating in Proceedings of the European Court of Human Rights

x.c.hr.al.1997

Protocole relatif à la Charte africaine des droits de l'homme et des peuples portant création d'une Cour africaine des droits de l'homme et des peuples

Protocol to the African Charter on Human and Peoples' Rights on the Establishment of an African Court of Human and Peoples' rights

x.c.hr.ce.2000

Protocole no. 12 à la Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales

Protocol No. 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms