Guide de l'annotation

1. Le système de balisage

Chaque forme qui apparait dans le corpus (token) est entourée par une balise <w></w> (word), dont le premier élément a été enrichi des trois étiquettes suivantes:

1. *lemma* avec un lemme en tant que forme de base
2. *pos* (part of speech) indiquant la catégorie grammaticale
3. *msd* (morphosyntactic description)[1]

Exemple : dans <w lemma="faire" pos="VER" msd="pres:1:pl">faisons</w>, la forme *faisons* est associée avec le lemme *faire*. De plus, elle est classée comme verbe (VER) au présent à la première personne du pluriel.
Remarque : puisque *msd* contient plusieurs valeurs, les requêtes des valeurs isolées nécessitent des jokers, p.ex. [msd="pres.*"] pour les formes du présent, [msd=".*1.*"] pour la première personne ou [msd=".*1:sg"] pour la première personne au singulier.

2. Les balises *pos* avec leurs valeurs *msd*

ADJ : adjectif ; valeurs msd : masc, femi, sg, pl
Exemple : <w lemma="actuel" pos="ADJ" msd="femi:sg">actuelle</w>

ADV : adverbe ; sans msd
Exemple : <w lemma="mieux" pos="ADV">mieux</w>

DET:ART : article, à l'exception des amalgames de préposition et article défini (v. ci-dessous sous PRP~DET) ; valeurs msd : masc, femi, sg, pl
Exemple : <w lemma="le" pos="DET:ART" msd="masc:pl">les</w>
Remarques : l'élément *l'* dans *l'on* a été traité comme un article défini séparé du pronom indéfini *on*. Les requêtes concernant l'article défini doivent tenir compte des amalgames avec les prépositions *à* et *de*. P.ex., ainsi seulement des requêtes comme [lemma="le" | lemma="de~le" | lemma="à~le"] saisissent toutes les occurrences de l'article défini. Les formes de l'article partitif sont codés comme PRP~DET.

1 Cf. le *STTS-large tagset*, https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf.

DET:POS : possessif en fonction de déterminant (au contraire de PRO :POS, v. ci-dessous) ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="notre" pos="DET:POS" msd="femi:sg">notre</w>

INT : interjection ; sans valeurs msd

Exemple : <w lemma="ah" pos="INT">Ah</w>

KON : conjonction ; sans valeurs msd

Exemple : <w lemma="mais" pos="KON">mais</w>

NAM:DIV : noms propres de divers types, c'est-à-dire noms à l'exception de NAM:GEO, NAM:PAR et NAM:PER (v. ci-dessous) ; sans valeurs msd

Exemple : <w lemma="Concorde" pos="NAM">Concorde</w>

Remarque : les acronymes qui représentent des noms propres (à l'exception de noms géographiques, de partis et groupes parlementaires et de personnes) ont été balisés avec pos="NAM:DIV", p.ex. <w lemma="ONU" pos="NAM:DIV">ONU</w>.

NAM:GEO : noms d'unités géographiques ; sans valeurs msd

Exemple : <w lemma="Bretagne" pos="NAM:GEO">Bretagne</w>

NAM:PAR : noms de partis et groupes parlementaires à l'Assemblée Nationale

Exemple : p.ex. <w lemma="LaREM" pos="NAM:ORG">LaREM</w>

Remarque : les noms de partis composés ont été balisés comme des unités fixes, p.ex. <w lemma="Les Républicains" pos="NAM:PAR">Les Républicains</w> (ainsi, ce nom de parti ne dérange pas les requêtes concernant l'article défini, entre autres).

NAM:PER : nom de personne ; sans valeurs msd

Exemple : *Marc Le Fur* comme <w lemma="Marc" pos="NAM:PER">Marc</w>
<w lemma="Le Fur" pos="NAM:PER">Le Fur</w>

Remarques : l'exemple montre que le prénom et le nom d'une personne sont pourvus de deux étiquettes séparées. Les noms composés, au contraire, ont été lemmatisés comme des unités (avec l'espace blanc faisant partie de la forme). Ainsi, les articles figés dans les noms de personne ne dérangent pas dans les requêtes concernant l'article. Les noms de personne qui figurent dans les noms des lois ont été classés comme NAM:DIV, p.ex. le nom de Michel Sapin dans *loi Sapin*.

NOM : substantif ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="couronne" pos="NOM" msd="femi:sg">couronne</w>

Remarques : pour les composés *ad hoc* avec des éléments similaires à des préfixes on a suivi le *Trésor*

de la Langue française informatisé (TLFi, http://atilf.atilf.fr/tlf.htm). Ainsi, p.ex. *mini-* est qualifié d'„élém. formant", *demi-* de „préf[ixe]". Pour cela, *demi-heure* et *mini-ports* ont été lemmatisés comme *demi-heure* et *mini-port*, respectivement. Pour les noms de métier féminins, les lemmes manquants ont été ajoutés manuellement, p.ex. *députée*, *présidente*. Les acronymes ont été balisés le plus précisément possible, c'est-à-dire avec les valeurs pos et msd, p.ex. : <w lemma="ASE" pos="NOM" msd="femi:sg">ASE</w>. L'abréviation n°. a été lemmatisé comme lemma="numéro". Les composés chimiques CO2 et CO ont été traité comme abréviations avec les lemmes CO2 et CO, respectivement.

NUM : numéral à l'exception des ordinaux (v. ci-dessous) ; sans valeurs msd

Exemple : <w lemma="vingt" pos="NUM">vingt</w>

Remarque : les numéros des articles de loi ont été balisés de la manière suivante : <w lemma="article" pos="NOM" msd="masc:sg">article</w><seg type="UA">L. 114-2</seg>.

NUM:ORD : numéral ordinal ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="premier" pos="NUM:ORD" msd="masc:sg">premier</w>

PRO : pronoms à l'exception de PRO:DEM, PRO:DET, PRO:IND, PRO:PER, PRO:POS, PRO:REF, PRO:REL (v. ci-dessous) ; sans valeurs msd

Exemple : <w lemma="quoi" pos="PRO">Quoi</w>

PRO:DEM : pronoms démonstratifs ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="celui" pos="PRO:DEM" msd="masc:sg">celui</w>

Remarques : cette catégorie comprend aussi les adjectifs démonstratifs, p.ex. *ce* dans *ce point* ou *telle* dans *une telle provocation*. Les pronoms neutres *ce* et *cela/ça* n'ont pas de valeurs msd.

PRO:IND : pronoms indéfinis ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="plusieurs" pos="PRO:IND" msd="masc:pl">plusieurs</w>

Remarque : cette catégorie comprend aussi les adjectifs indéfinis, p.ex. *toute* dans *toute mission*.

PRO:PER : pronoms personnels ; valeurs msd : 1, 2, 3,[2] masc, femi, sg, pl

Exemple : <w lemma="se" pos="PRO:PER" msd="3:masc:pl">se</w>

PRO:POS : pronoms possessifs ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="sien" pos="PRO:POS" msd="femi:sg">sienne</w>

2 Les nombres indiquent la personne.

PRO:REL : valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="dont" pos="PRO:REL">dont</w>

Remarque : le jeu d'étiquettes du TreeTagger utilisé ne connait pas la catégorie des interrogatifs. Pour cela, les formes *quel*, *quelle*, *quels*, *quelles* et *qui* ont été balisées comme PRO:REL.

PRP : préposition, à l'exception des amalgames de préposition et article défini (v. ci-dessous) ; sans valeurs msd

Exemple : <w lemma="en" pos="PRP">en</w>

Remarque : les requêtes concernant les prépositions doivent tenir compte des amalgames avec l'article défini. P.ex., seulement la requête [lemma="de" | lemma="de~le"] saisit toutes les occurrences de la préposition *de* et la requête [pos="PRP.*"] saisit toutes les prépositions.

PRP~DET : amalgame de prépos. et article défini (*au*, *aux*, *du* ou *des*) ; valeurs msd : masc, femi, sg, pl

Exemple : <w lemma="de~le" pos="PRP~DET" msd="masc:sg">du</w>

Remarque : Les formes de l'article partitif sont codés comme PRP~DET.

SYM : symboles ; sans valeurs msd

Exemple : <w lemma="%" pos="SYM">%</w>

SYM:NUM : numéros de lois et amendements ; sans valeurs msd

Exemple : <w lemma="47" pos="SYM:NUM">47</w> (dans amendement no. 47)

VER : verbe ; valeurs msd : cond (conditionnel), futu (futur), impe (impératif), impf (imparfait), infi (infinitif), pper (participe passé), ppre (participe présent), pres (présent), simp (passé simple), subi (subjonctif imparfait), subp (subjonctif présent), 1, 2, 3,[3] masc, femi (participes), sg, pl

Exemple : <w lemma="être" pos="VER" msd="pres:3:sg">est</w>

3. Les informations concernant les locuteurs et locutrices

Les énoncés des débats dans les comptes rendus sont mis en relation avec les locuteurs et locutrices. Cela permet de créer des sous-corpus et des partitions en fonction de leur âge, sexe, groupe parlementaire, parti ou circonscription électorale. Ces valeurs se trouvent groupées sous l'attribut « sp ».

3 Les nombres indiquent la personne.