# Guide de préparation de corpus pour soumission à SATO

version 1.2 (janvier 2010)

## I. PRINCIPES ET DÉFINITIONS

### Qu'est-ce qu'un corpus?

**Définition.** On utilise généralement le terme de *corpus* pour désigner un ensemble raisonné de textes. Dans le contexte d'une analyse de texte assistée par ordinateur, le corpus correspond aux données qu'on sélectionne et qu'on organise pour répondre aux fins particulières de la recherche.

Comme tout jeu de données, le corpus doit donc faire l'objet d'une organisation et d'une description dont on tirera profit lors de l'analyse. Plus encore, l'analyse elle-même aura pour effet d'ajouter aux données textuelles de départ des niveaux supérieurs de description destinés à en expliciter le fonctionnement par l'application de procédures sous gouverne de l'analyste.

Les données textuelles ne sont donc pas simplement un flux de caractères dans un fichier d'ordinateur. Ce sont des données déjà structurées en entrée de l'analyse qui produiront des données plus structurées en sortie. L'analyse sur corpus a un caractère itératif puisque la sortie d'une première analyse est susceptible de servir d'entrée à une seconde analyse sur des textes enrichis par divers niveaux d'annotation. Un logiciel d'analyse textuelle, tel SATO, prévoit donc une syntaxe de présentation des données qui permet de représenter des couches d'annotation de diverses natures et de complexité grandissante au fur et à mesure de l'évolution des analyses. L'ajout de catégories à des mots ou des segments de texte est une des manières d'annoter un corpus.

Ce *Guide de préparation des corpus*, vise à fournir un certain nombre de recommandations pour la mise en forme initiale des corpus, en amont de l'analyse informatisée. En même temps, nous tenterons d'identifier, en aval, des opérations supplémentaires qui feront partie des premières analyses SATO que l'usager pourrait vouloir déployer, s'il en a besoin, afin de produire des versions enrichies de son corpus. En d'autres mots, la méthode proposée ici est d'y aller par étape, en prenant, à chaque étape, les décisions jugées nécessaires pour l'interprétation des données en fonction des objectifs de recherche de l'analyste.

Dès lors, tout comme pour d'autres logiciels d'analyse de textes, il est nécessaire de fournir à SATO des documents en format texte brut. En SATO, un corpus se présente donc comme une série de documents précédée d'un entête qui documente l'ensemble du corpus.

### Qu'est-ce qu'un fichier en format texte?

L'analyse d'un corpus de textes à l'aide de l'ordinateur implique que l'on puisse disposer de documents en format numérique, plus précisément de fichiers dans lesquels les mots sont représentés sous forme de caractères.

**Définition.** Un document texte est un fichier de caractères que l'ordinateur est capable d'ouvrir au moyen d'une application permettant d'en effectuer la lecture et/ou l'édition.

L'édition d'un document texte consiste, par exemple, en l'action d'effectuer un changement d'orthographe, de faire un copier-coller ou encore, de rajouter des phrases. Ainsi, les textes disponibles au format Word, ayant par exemple l'extension *.doc* ou *.rtf*, ainsi que les fichiers ayant l'extension *.txt* sont des documents textes. Certains fichiers *.pdf* sont également des documents textuels. Ce n'est pas le cas de tous les documents PDF.

**Attention.** Certains documents peuvent afficher du texte, mais ils n'en permettent pas l'édition. C'est le cas de nombreux fichiers PDF. En effet, certains documents peuvent afficher de l'information que notre œil interprète comme étant du texte, mais qui sont en fait des images que l'ordinateur ne peut éditer ou transformer sans opérations additionnelles. Pour être transformés en documents textes, ces images doivent faire l'objet d'une conversion en caractères au moyen d'un logiciel de reconnaissance optique de caractères. Certains logiciels permettent de numériser directement un fichier d'images en format PDF. Sinon, on doit imprimer le document PDF et le numériser comme s'il s'agissait d'un document papier.

Parmi les documents en format texte, on distinguera ceux qui contiennent des codes internes qui ne peuvent être lus que par des logiciels spécialisés, par exemple les fichiers en format Word, des documents en format texte brut peuvent être lus, affichés et imprimées par tous les logiciels qui manipulent du texte. C'est ce format qui est supporté par SATO.

**Qu'est-ce que le balisage?**

Si le corpus n'est pas qu'un simple flux de caractères, c'est parce que, dans la séquence des caractères, on retrouve des balises qui servent à donner de l'information sur les caractères qui suivent la balise.

**Définition.** Une balise est une séquence de caractères servant à donner des informations sur le contenu du texte et de ses diverses parties. Les balises permettent donc d'annoter le corpus pour en décrire la structure et l'enrichir de différents points de vue. Une balise est généralement introduite par des *métacaractères*, c'est-à-dire des caractères qui seront interprétés non pas comme du contenu textuel, mais comme des caractères spéciaux servant de délimiteurs de balises.

Le balisage est donc le procédé technique qui permet de codifier un corpus pour le décrire et l'annoter.

Le stylage qu'on utilise dans un traitement de texte afin de mettre un titre en gras, un mot étranger en italique, etc. est une forme de balisage, puisque, dans le fichier interne du traitement de texte, on doit faire la distinction entre des attributs de présentation et le contenu présenté. Ce balisage d'édition a cependant l'inconvénient de confondre le détail du rendu visuel avec la raison qui a poussé le rédacteur à choisir ce rendu visuel.

Par exemple, si on utilise l'italique pour marquer un mot étranger dans un article, mais aussi pour marquer le titre d'un ouvrage dans la bibliographie, on obtient un balisage qui peut être interprété par un lecteur humain habitué à ce style d'écrits. Mais pour l'ordinateur, l'interprétation de l'usage de l'italique est beaucoup plus complexe. Dans notre exemple, l'ordinateur aura besoin d'une balise supplémentaire introduisant une variable dont les valeurs possibles correspondraient aux divisions structurelles de l'article (Titre, Sous-titre, Résumé, Bibliographie, Notes, etc.). Dans ce cas, on pourra toujours demander au programme d'*interpréter* la variable de présentation, avec sa valeur italique dans notre exemple, en fonction de la structure de l'article, *Bibliographie* par exemple. Cet exemple montre que les balises permettent d'introduire des variables sur des objets textuels à différents niveaux de granularité de l'information : corpus en entier, documents dans le corpus, sections dans le document, phrases dans la section, mots dans la phrase, etc.

Comme dans l'analyse des données quantitatives, on distingue les variables par leur type et leur domaine d'application. Par exemple, les variables *division* et *présentation,* dans notre exemple, sont de type modales ou catégorielles, c'est-à-dire que leurs valeurs sont des *catégories* (on dit aussi de *modalités*) appartenant à un ensemble fini de symboles. On notera aussi que la variable *division* concerne davantage la structure du document en distinguant les parties du texte, alors que la variable *présentation* concerne davantage les mots et les caractères.

Pour schématiser, on pourrait dire que les variables, et les balises qui servent à les coder, se distinguent selon la dimension de l'empan textuel annoté. Au niveau *macro*, les variables pourraient être qualifiées d'*externes,* en ce qu'elles s'appliquent à l'ensemble d'un document et qu'elles visent à en décrire les conditions de production : lieu, date, auteur, genre, référence bibliographique, etc. Ces données font souvent partie des métadonnées documentaires, un peu à la manière d'une fiche dans un catalogue de bibliothèque.

Le deuxième niveau d'annotation, que l'on pourrait qualifier d'intermédiaire (*méso*) s'intéresse à la structure formelle du document en termes de parties. Dans le cas d'une transcription d'entrevue par exemple, il pourrait s'agir de l'identification des locuteurs, dans le tour de parole, en plus de la division en questions pour une entrevue semi-dirigée. Pour un article de journal, la structure du document pourrait comprendre les parties suivantes : localisation dans le journal, titre, sous-titres, date de publication, nom du journaliste, etc. Ces variables permettent donc de rendre compte du caractère composite d'un document et d'en expliciter la structure.

Enfin, on pourrait parler d'un niveau *micro* pour décrire, par exemple, diverses unités linguistiques et terminologiques : catégories grammaticales, catégories socio-sémantiques noms de lieu, acronymes, locutions, etc.

**Quelle est la convention utilisée par SATO pour les balises?**

Il existe de multiples syntaxes concrètes pour introduire des balises. Il existe même des langages qui ont pour fonction de définir des protocoles de balisage. Le plus connu est XML.

**Définition.** XML est l'acronyme de l'expression anglaise *Extensible Markup Language*. Il s'agit en fait d'un métalangage qui permet de définir une variété de langages de balisage ayant tous en commun une même syntaxe de base. XML est une recommandation du *World Wide Web Consortium's*.

Il existe aussi des organismes qui se donnent pour mission de suggérer des noms de balises en XML pour codifier des corpus de textes. C'est le cas, en particulier, du *Text Encoding Initiative*.

**Définition.** TEI est le sigle du *Text Encoding Initiative Consortium*. Cet organisme est surtout connu pour son *Guide* contenant un ensemble de recommandations pour le marquage de documents textuels.

Un ensemble restreint de balises TEI a été retenu comme langage pivot permettant d'échanger des corpus entre plusieurs logiciels de textométrie dont les auteurs se retrouvent au sein du réseau ATONET (*Réseau pour l'échange de ressources et de méthodologies en analyse de texte par ordinateur*).

L'interface à SATO comprend un programme, basé sur ce langage pivot XML-TEI, permettant de convertir des corpus d'un format logiciel à l'autre.

Le logiciel SATO existait déjà avant l'arrivée de XML. Il propose donc sa propre syntaxe de balisage conçue pour être simple à utiliser par des non-spécialistes. Par ailleurs, dans le cadre du projet ATONET (http://www.atonet.net/), plusieurs développeurs de logiciels de Textométrie ont convenu d'un format pivot utilisant un nombre minimal de balises XML-TEI permettant de convertir un corpus balisé pour un logiciel donné, SATO par exemple, vers un autre logiciel en passant par ce formalisme pivot.

Dans SATO, les variables de balisage sont appelées des *propriétés*.

Voici un exemple de balisage avec la syntaxe SATO. Cet exemple utilise une propriété *partie* dont l'objectif est de distinguer des parties d'un texte : *titre*, *sous-titre*, *texte* etc. Pour faciliter la lecture de l'exemple, les balises sont en gras.

```
*partie=titre Le monde est à cours de pétrole!
*partie=sous-titre La communauté scientifique est aux abois
*partie=texte Depuis la Seconde Guerre mondiale, nous avons accru notre
dépendance aux produits dérivés du pétrole de manière radicale. En plus
de dépendre du pétrole pour nos déplacements, nous sommes entourés de
produits dont le pétrole est une des composantes essentielles.
*partie="nil"- 30 -
```

Dans cet exemple, la première ligne est précédée d'une balise de propriété SATO. La propriété (variable SATO) *partie* reçoit la valeur *titre* indiquant que les mots qui suivent font partie d'un titre. La deuxième ligne commence par une balise qui affecte la valeur *sous-titre* à la propriété *partie*, indiquant que le texte qui suit est un sous-titre. La troisième ligne débute par une balise indiquant que les mots qui suivent feront partie du corps du texte (valeur *texte)*. Finalement, pour identifier la convention indiquant la fin d'un communiqué (– 30 –) on a utilisé la valeur par défaut *nil* qui représente l'absence de catégorie. Ainsi balisé, on pourra, en analyse, utiliser la valeur de la propriété *partie* pour concentrer l'analyse sur un ou plusieurs types de section.

De façon plus formelle, voici comment se présentent les balises de propriété dans un corpus en format SATO.

1. Toutes les balises de SATO commencent par une étoile « * ». Ce signe est celui défini par défaut. Au besoin on peut choisir un signe alternatif comme *métacaractère* dans la configuration du corpus.

**Attention.** Il est important de s'assurer qu'outre les balises, l'usage des étoiles (ou tout autre signe choisi pour introduire une balise) dans le corpus soit contrôlé. Le corps du texte peut contenir des étoiles à la seule condition qu'elles soient précédées d'un *caractère de citation* (on dit aussi *caractère d'échappement*). L'étoile qui suit immédiatement le caractère de citation sera alors considéré comme un caractère du texte plutôt que comme un métacaractère. Le caractère de citation, s'il n'est pas modifié par une commande de déclaration en tête du corpus, est \ (barre oblique inverse). On utilisera la syntaxe suivante pour citer une barre oblique : \\

2. Collé à l'étoile, se trouve une chaîne de caractère qui est le nom de la propriété. Un nom de propriété est un identificateur standard. Le nom de la propriété n'est pas sensible à la casse.

**Définition.** Dans la syntaxe de SATO, un identificateur standard commence par une lettre qui peut être suivie de lettres ou de chiffres. On peut aussi utiliser les caractères – et _ à l'intérieur d'un identificateur standard.

3. Collé directement à la droite du nom de la propriété, se trouve le signe égal « = ». Le signe égal introduit la ou les valeurs que va prendre la propriété dans ce contexte.

4. Directement collées à droite du signe égal, on retrouve la ou les valeurs attribuées à la propriété. Dans l'exemple, les valeurs sont des symboles qui suivent la syntaxe des identificateurs standards. Une valeur de propriété qui ne suit pas cette syntaxe doit être encadrée par des guillemets anglais ("). Un identificateur standard peut aussi être mis entre guillemets. Dans la troisième ligne de l'exemple, la valeur de la propriété est entre guillemets anglais (**"nil"**) pour éviter de la démarquer du contenu textuel (– 30 –) qui est collé sur la balise. Donc, on peut formaliser la syntaxe d'une balise ainsi :

***propriété=valeur**

**Attention.** On notera que toutes les composantes d'une balise SATO sont collées!

Dans SATO, une balise qui n'est pas collée sur le mot qui la précède s'applique à tous les mots qui la suivent jusqu'à ce qu'une balise introduisant une autre valeur de la même propriété soit entrée. C'est un peu comme si nous avions pris un marqueur fluorescent et qu'on le laissait courir jusqu'à temps de changer pour un marqueur d'une autre couleur.

**Attention.** Une nouvelle valeur de propriété annule la précédente si, et seulement si, il s'agit de la même propriété.

**La syntaxe des balises de propriété SATO peut-elle être plus élaborée?**

En effet, dans certains cas, il peut être utile d'avoir une syntaxe plus élaborée. Il est rare qu'on balise directement un corpus *à la main*. Cependant, quand on exporte un corpus qui a été enrichi en analyse, on peut se retrouver avec des cas de balisage qui nécessitent une syntaxe élaborée. On la donne donc ici à titre informatif.

1. Ainsi, le signe d'égalité (=) qui suit le nom de la propriété peut être précédé de + ou de - pour indiquer que la valeur de propriété doit être ajoutée ou soustraite de la valeur courante. Voici un exemple.

```
*style=gras Ce texte est en gras, ce qui empêche pas d'avoir de l'italique*style+="italique" en plus!
```

Dans cet exemple le mot *italique* a reçu la valeur « gras » **et** la valeur « italique ».

2. Une balise de propriété collée à la droite d'un mot n'affecte que ce mot. C'est ce que l'on appelle une affectation locale de valeur de propriété. Dans le cas contraire, on parle d'affectation globale. Dans l'exemple qui précède, l'ajout de l'italique à la phrase en gras ne concerne que le mot *italique* lui-même. Voilà pourquoi il est directement collé sur le mot. Quelques fois, cette interprétation est indésirable. C'est le cas, par exemple, si on voulait mettre en italique toute l'expression *italique en plus* mais sans inclure le « l' ». Le truc serait alors d'introduire le pseudo-espace *_ juste après le « l' ». Voici l'exemple.

```
*style=gras Ce texte est en gras, ce qui empêche pas d'avoir de l'*_*style+="italique"italique en plus*_*style-="italique"!
```

Cet exemple indique que la valeur *italique* s'ajoute (**+=**) à la valeur gras pour tous les mots qui suivent jusqu'à ce qu'on l'enlève (-=) ! Le pseudo espace *_ permet simplement de dire à SATO d'interpréter la balise comme une affectation globale même s'il n'y a pas de vrais espaces séparant la balise du mot qui précède.

3. Comme indiqué avec la propriété *style* de l'exemple précédent, une propriété dont les valeurs sont des symboles désignant des catégories est susceptible de recevoir plus d'une catégorie. On peut affecter directement un ensemble de symboles à une propriété SATO. Dans ce cas, les symboles sont séparés par des virgules et l'ensemble est entre parenthèses. Voici un exemple.

```
Le grizzli*classification=(mammifère,omnivore) est un animal solitaire
mais le caribou*catégorie=(mammifère,herbivore) est un animal grégaire.
```

**Attention.** Dans un ensemble de symboles, on ne doit pas mettre d'espaces.

**Est-ce qu'il y a plusieurs sortes de propriétés dans SATO?**

Dans SATO, on distingue diverses classes de propriétés selon deux critères :

1) le type de valeurs qu'elles peuvent contenir et
2) leur portée.

Une propriété a une portée textuelle si elle concerne chacun des mots en contexte. C'est dire qu'elle peut servir à marquer des segments de texte dont la longueur minimale est d'un mot.

Une propriété a une portée lexicale si elle concerne la forme lexicale, c'est-à-dire la graphie normalisée du mot, telle que répertoriée dans le lexique du corpus. Une propriété lexicale, utilisée dans la préparation d'un corpus à soumettre à SATO, permettra de distinguer des graphies identiques mais ayant des valeurs différentes de propriété lexicale. On peut, par exemple, utiliser une propriété lexicale pour lever une ambiguïté. Dans l'exemple suivant, on utilise une propriété que l'on a nommé *sens* pour qualifier des mots dont on cherche à retirer l'ambiguïté.

```
Quand je vais à la pêche, je n'oublie jamais d'emporter mes
vers*sens="animal". Lorsque je les accroche à l'hameçon, je tâche de
les installer vers*sens="direction" le centre.
```

Ici, on voit qu'on a retiré l'ambiguïté du mot *vers* qui signifie à la fois une direction et un animal. Identifié ainsi, le mot *vers* apparaîtra deux fois dans le lexique du corpus SATO alors que le mot *je* (qui apparaît trois fois dans le texte) n'apparaîtra qu'une seule fois dans le lexique.

Voilà pour la *portée* de la propriété. Voici maintenant ce que l'on entend par le *type* de la propriété.

Une propriété est dite *symbolique*, si elle peut prendre comme valeurs un ou plusieurs

symboles (ou catégories, ou modalités) parmi un ensemble défini de symboles. Sa valeur par défaut est le symbole *nil* qui désigne un ensemble vide, c'est-à-dire aucun symbole. Les propriétés de ce type servent généralement à catégoriser les mots.

Une propriété est dite *entière* si les valeurs qu'on peut lui affecter sont des nombres entiers positifs ou zéro. Sa valeur par défaut est 0. Les propriétés de ce type servent généralement à compter des mots.

Une propriété est dite en format *libre* si elle peut prendre comme valeur n'importe quelle chaîne de caractères ne dépassant pas 200 caractères environ. Sa valeur par défaut est la chaîne vide. Ce type de propriété permet de gérer des annotations libres sur un mot en contexte (*occurrence*) ou sur une forme lexicale.

**À quoi sert la propriété page?**

SATO dispose d'une propriété prédéfinie destinée à contenir la référence de pagination de chaque mot du corpus. Cette propriété contient une valeur structurée en quatre champs :

1. nom du document;
2. numéro de la page dans le document;
3. numéro de la ligne dans la page;
4. numéro du mot dans la ligne.

Cette propriété est gérée automatiquement lors de la soumission d'un corpus.

1. Le compteur de numéros de mots est incrémenté de 1 à chaque nouveau mot dans la ligne.
2. Le compteur de lignes est incrémenté de 1 à chaque nouvelle ligne de texte marquée par une fin de ligne. Un saut de ligne automatique est aussi généré lorsqu'on atteint un nombre maximum de mots par ligne. Il est à noter que les lignes vides ne sont pas comptées. Elles sont cependant reconnues comme délimiteurs de paragraphes.
3. Le compteur de pages est incrémenté à chaque fois qu'on atteint un nombre maximum de lignes par pages. Mais, on peut aussi provoquer un saut de page par une marque dans le corpus. Il peut s'agir du caractère de saut de page (*form feed*) ou d'une balise explicite de pagination.
4. Le nom du document n'est jamais changé automatiquement. Mais, en l'absence de toute balise de pagination, le nom implicite du document sera *Doc1*.

La référence de pagination en SATO correspond à la valeur de la propriété page. Voici quelques exemples.

**Référence 1**

***page=ton_livre** Texte complet du document ton_livre

La première référence (***page=ton_livre**) se limite au nom du document, sous la forme d'un identificateur standard. SATO assumera que le document commence au premier mot de la première ligne de la première page du document. À partir de là, la pagination automatique

sera utilisée à l'intérieur du document.

**Référence 2**

```
*page=mon_livre/5/2/3
Extrait du document de mon_livre en page 5, à partir du troisième mot
de la ligne 2...
```

La deuxième référence (***page=mon_livre/5/2/3**) est très précise et cite un extrait de *mon_livre* commençant au troisième mot de la deuxième ligne de la cinquième page du document *mon_livre!* À partir de là, la pagination automatique sera utilisée à l'intérieur du document.

**Référence 3**

```
*page=/7 Suite de mon_livre au début de la page 7
```

La troisième référence (***page=/7**) indique que l'extrait de *mon_livre* se poursuit au début de la septième page (on présume ici que cette balise suit celle de l'exemple 2). À partir de là, la pagination automatique sera utilisée à l'intérieur du document.

**Référence 4**

***page=@critique_de_paul.txt**

Enfin, la quatrième référence (***page=@critique_de_paul.txt**) introduit un nouveau document dont le contenu se trouve sur un fichier séparé envoyé sur le serveur. Le nom du fichier est *critique_de_paul.txt* et le nom du document dans SATO sera *critique_de_paul*. Il est aussi possible d'avoir un nom de document différent du nom du fichier, par exemple *paul*, en utilisant la syntaxe suivante : ***page=paul@critique_de_paul.txt**. À partir de là, la pagination automatique sera utilisée à l'intérieur du document.

Puisque la référence de pagination se présente sous la forme d'une propriété SATO, il sera possible, lors de l'analyse du corpus, d'inclure cette propriété dans un filtre pour sélectionner une partie du corpus en fonction de sa pagination. Ainsi, un choix judicieux des noms de documents permettrait de coder simplement des informations concernant l'ensemble du document. On trouvera, dans la section *Méthodes et procédures*, des exemples de noms informatifs.

**Est-ce qu'on peut avoir un corpus en plusieurs langues?**

SATO est en mesure de traiter des textes multilingues codés selon divers alphabets. La propriété prédéfinie *alphabet* permet de passer d'une langue à l'autre. Voici un exemple.

```
*alphabet=fr Ce corpus est en français mais on trouve aussi des
citations *alphabet=en in english! *alphabet=fr dans la bibliographie.
```

La définition des divers alphabets utilisés dans le corpus fera partie de l'entête du corpus, tel qu'expliqué dans la section *Méthodes et procédures*.

**Comment SATO considère-t-il les lettres en majuscules?**

SATO, lorsqu'on lui soumet un texte, décode les balises de propriété et repère les mots en fonction de l'alphabet en vigueur. La forme graphique de chacun des mots est conservée dans le lexique du corpus. Le sens des mots est généralement indépendant de leur mise en forme utilisant, en particulier, la majuscule après un point de phrase. Voilà pourquoi, SATO normalisera la graphie en ne conservant au lexique que la forme du mot en caractères minuscules. Cela dit, l'information sur la forme du mot n'est pas perdue. Elle est conservée comme valeur de la propriété *Édition*, une propriété prédéfinie dans SATO. Cette propriété conserve, pour chacune des occurrences des formes lexicales, des indications indiquant comment la forme normalisée en minuscules doit être présentée lors de l'écriture du texte.

D'autres attributs d'édition sont automatiquement dépistés par SATO et notés comme valeur de la propriété *Édition* : *maj* (pour majuscule de début de mot), *cap* (pour mot tout en lettres capitales), *par* (pour un mot en début de paragraphe), *collé* (pour un mot collé sur le suivant comme le *l'* dans *l'école*), *sic* (pour indiquer que la graphie du mot ne suit pas les règles de l'alphabet), etc. Comme pour les autres propriétés, la propriété *Édition* pourra être utilisée lors de l'analyse pour désigner des mots en majuscules, en début de paragraphe, etc.

Il arrive quelques fois que la majuscule doive faire partie de la forme lexicale, par exemple pour distinguer le nom commun *pierre* du nom propre *Pierre*. Pour conserver la majuscule au lexique dans le cas du nom propre, il faut faire précéder la majuscule du métacaractère de citation, soit la barre oblique inverse, si le caractère de citation n'a pas été redéfini dans l'entête du corpus. Dans notre exemple, cela donnerait ceci.

```
\Pierre possède une pierre précieuse.
```

**Attention.** Il n'est pas recommandé de citer *à la main* les noms propres puisque des procédures SATO permettent de faciliter cette tâche si on juge nécessaire, pour l'interprétation, de lever les cas d'ambigüité potentielle entre un nom propre et un mot grammatical.

**Comment SATO peut-il reconnaître les mots ayant une forme complexe : mots composés, sigles, expressions, etc.?**

SATO se base sur les règles d'écriture alphabétique de la langue pour découper le texte en mots. Ces règles peuvent être insuffisantes parce que le code alphabétique est souvent ambigu. C'est le cas du trait d'union qui est utilisé à la fois pour les mots composés et l'inversion du pronom et du verbe dans la forme interrogative comme dans *aimes-tu*. C'est le cas aussi de certains noms propres, sigles et expressions composés de mots séparés par des espaces, comme dans *assemblée nationale*. Pour forcer SATO à considérer ces chaînes de caractères comme des entrées dans le lexique du corpus, on peut les encadrer par les balises **(* et **)* d'ouverture et de fermeture de mot. Voici des exemples.

```
La question nationale a été soulevée à l'*(assemblée nationale*). As-tu
vu ce *(m'as-tu vu*)?
```

Les expressions *assemblée nationale* et *m'as-tu vu,* ainsi balisées, seront inscrites telles quelles dans le lexique du corpus.

**Attention.** Comme des procédures SATO sont prévues pour faciliter le balisage des mots-composés, des expressions et des abréviations, on ne conseille pas de le faire *à la main* sur le corpus en format texte.

**Comment garder des parties de texte à des fins de documentation et empêcher SATO de les analyser?**

Il est possible de mettre des parties de texte en commentaire afin d'empêcher SATO d'en tenir compte lors de l'analyse. C'est le cas, par exemple, de tableaux statistiques à l'intérieur d'un document. Pour indiquer à SATO d'ignorer une partie dans un texte, il faut encadrer cette partie par une balise de commentaire. Cette balise est introduite par une étoile (ou le métacaractère défini dans l'entête du corpus comme caractère de propriété). Ensuite, on colle à l'étoile une accolade gauche « { ». On fini le commentaire en rajoutant un « } ». Voici un exemple.

```
*{Donc, dans SATO, tout le texte qui se trouve entre ces signes est
rejeté de l'analyse, mais il est gardé en mémoire et sera affiché au
besoin}
```

**Attention.** Sauf exception, par exemple la suppression d'informations confidentielles, il est préférable de mettre en commentaire les parties indésirables du texte plutôt que de les supprimer. Autant que possible, en effet, on essaie de conserver l'intégrité documentaire d'un texte afin de pouvoir le réutiliser dans d'autres contextes d'analyse et afin d'avoir une trace explicite de toute modification apportée à des fins particulières.

## 2. MÉTHODES ET PROCÉDURES

## 2.1. Collecte d'un corpus à partir de différents types de documents

### Que faire dans le cas de documents oraux enregistrés?

La manière la plus simple de recueillir des discours oraux enregistrés pour l'analyse dans SATO est d'effectuer une retranscription.

**Définition.** La retranscription consiste à écrire dans un document texte le contenu des échanges oraux de même que les informations permettant de comprendre le contexte des échanges. La complexité du protocole de transcription varie selon que l'on s'intéresse à l'oralité en tant que telle, ou au contenu des échanges.

L'analyse linguistique de la parole implique des protocoles élaborés de transcription phonétique et alphabétique accompagnés, dans le cas des enregistrements vidéos, du codage du langage gestuel. Il existe des logiciels spécialisés permettant de faciliter ces retranscriptions sur plusieurs lignes synchronisées, à la manière d'une partition musicale.

Dans le cas où seul le contenu des échanges nous intéresse, on procèdera directement à une transcription alphabétique accompagnée d'annotations permettant de rendre compte du contexte des échanges.

Concernant la transcription de la parole, on doit d'abord convenir de ce que l'on fait des phénomènes typiques de l'oralité : reprises, hésitations, silences, soupirs et autres expressions non grammaticales. En général, la règle de conduite est tributaire des objectifs de la recherche. Dans un récit de vie, par exemple, il peut être intéressant de conserver les marques de l'émotion. Comme elles peuvent être nombreuses et difficiles à mettre en mots, une façon de faire est de fournir à la personne chargée de la transcription une grille permettant de traduire ces expressions par des codes distincts des mots de la langue, de telle sorte qu'ils puissent en être facilement distingués. Par exemple, plutôt que d'inscrire *(*rire), *(silence prolongé) (euhhh)* dans le texte, on écrira *_rire*, *_silence_prolongé* *_hésitation_simple*, etc. Dans le lexique du texte, *_rire* sera ainsi comptabilisé de façon distincte du mot *rire* qui aurait été prononcé en tant que tel. Dans le cas, par exemple, d'une entrevue informelle avec un expert, on peut penser que ces traits de l'oralité ne contribueront pas à l'analyse et on n'en tiendra simplement pas compte dans la retranscription.

On doit aussi convenir du niveau de langue dans la transcription. Les règles de l'expression orale diffèrent de celles de l'écrit. Les mots dits peuvent s'éloigner de l'orthographe d'usage et des règles d'accord. Si on veut faire appel à des outils de traitement automatique de la langue, il peut être utile de corriger la langue. Sinon, il est préférable de s'en tenir à une transcription plus littérale tout en ne variant pas trop l'orthographe. C'est le cas des jurons et tics de langue (*tabernouche*, *tabarnouche*...). Si les variantes phonétiques ne sont pas requises pour l'analyse, il est préférable de maintenir la même orthographe tout au long de la transcription.

On pourrait aussi vouloir codifier des éléments non verbaux, surtout dans le cas des

enregistrements vidéos. Dans ce cas, le plus simple serait de décrire l'événement en mots en le distinguant des échanges eux-mêmes. Si on n'entend pas approfondir l'analyse de ces événements, on se contentera de les mettre en commentaire. Sinon, on pourra considérer ces descriptions comme des interventions d'un locuteur hors champ agissant, pour ainsi dire, à titre de narrateur. Ainsi, s'il y a lieu, on pourra, lors de l'analyse du corpus, se pencher spécifiquement sur le contenu de la narration pour contextualiser les échanges oraux.

Finalement, si c'est utile, on pourra ajouter des commentaires SATO à titre de repères qui ne feront pas partie de l'analyse, mais qui pourront accompagner le texte. C'est le cas, par exemple, de la référence temporelle à l'enregistrement pour un retour à l'enregistrement, si nécessaire.

L'exemple suivant illustre un balisage élaboré de retranscription d'entrevues. On utilisera une propriété *locuteur* pour distinguer les interventions de l'interviewer de celles du répondant. Comme on est susceptible d'avoir un corpus avec plusieurs entrevues impliquant plus d'un répondant et peut-être plus d'un interviewer, la valeur de la propriété *locuteur* sera utilisée pour coder le profil de l'intervenant, par exemple L1 pour l'interviewer 1 et Rf22 pour désigner le répondant 22 qui est une femme. Marquées adéquatement, les questions et les réponses pourront ainsi être traitées séparément. De plus, en prévoyant des caractères pour coder le profil du répondant, il sera aussi facile, lors de l'analyse, de filtrer les noms des répondants pour ne sélectionner que les femmes, par exemple. Dans l'exemple, *T1* est utilisé comme valeur du locuteur pour baliser les commentaires du transcripteur-narrateur.

Comme il arrive souvent dans les entrevues dirigées ou semi-dirigées, on a prévu un plan d'entrevue abordant des thèmes particuliers. On utilisera une deuxième propriété, *section* dans l'exemple, qui distinguera les thématiques du plan d'entrevue.

***thématique=biographie**
***{**bande_1 400 s.**}**
***locuteur=I1** Qu'est-ce qui dans ton enfance t'a porté à t'intéresser à l'écologie?

***locuteur=T1** L'entrevue a été interrompue par la sonnerie du téléphone. Le répondant décide de ne pas répondre.

***locuteur=Rf22** Eh bien… J'avais… 8 ou 9 ans je crois. Je me rappelle qu'un ami m'avait invité chez eux. Ses parents faisaient partie d'une organisation écologique. Lorsque je suis arrivé chez eux, les parents discutaient d'une action qu'ils allaient mener contre une usine qui déversait des produits chimiques dans une rivière proche. […]

***thématique=opinion**
***{**bande_2 320 s.**}**
***locuteur=I1** Qu'est-ce que tu penses de ceux qui disent qu'on ne devrait pas impliquer les enfants en bas âge en politique?

***locuteur=Rf22** Je ne suis pas d'accord avec eux _soupir. _silence Je trouve que de toutes manières les enfants sont entourés de situations dans lesquelles il est important d'avoir une idée de ce que doit être la société. Et que de les impliquer le plus tôt possible est une des meilleures manières de les préparer à la vie en société.

Au Québec, plusieurs professionnels sont spécialisés dans la retranscription de corpus oraux. En 2008, il faut compter au moins 100$ pour la retranscription d'une heure de discours oral.

**Que faire dans le cas de documents imprimés sur papier?**

La manière la plus simple de recueillir des discours écrits sur papier pour l'analyse dans SATO est d'effectuer une numérisation.

Le numériseur (*scanner)* est le périphérique d'ordinateur permettant la numérisation d'images. Ainsi, chaque page-papier du corpus soumise au numériseur est traduite sous forme d'image électronique dans un fichier de l'ordinateur. Pour obtenir un fichier en mode texte, il faut transformer cette image en une suite de caractères. Pour ce faire, on doit soumettre l'image à un logiciel de reconnaissance optique des caractères, appelé communément OCR, pour *Optical Character Recognition* (reconnaissance optique des caractères). *OmniPage* est un des logiciels commerciaux permettant de réaliser cette tâche.

Il est nécessaire, d'une manière générale, d'envisager une moyenne de 2 à 4 minutes par page à numériser. Cette période comprend le passage mécanique de la lampe (première étape de la numérisation), ensuite la transformation des images obtenues en informations textuelles et, enfin, la confrontation de ces informations avec le dictionnaire interne du logiciel. Cette étape est d'autant plus longue et laborieuse que le document d'origine est de mauvaise qualité et que les termes utilisés appartiennent à un univers lexical distant du français commun. C'est particulièrement vrai si le texte est une retranscription d'un discours parlé comme des entretiens ou si le texte à une composante technique importante. La vérification manuelle du document produit peut donc, dans certains cas, être très longue.

**Pourquoi est-ce que je n'arrive pas à copier le texte de mon fichier PDF?**

Il existe au moins deux causes à l'impossibilité de copier de l'information texte à partir d'un fichier PDF.

1. La première cause possible réside dans le fait que ces fichiers peuvent avoir été protégés par leurs auteurs grâce à un mot de passe. Dans ce cas, deux solutions peuvent être envisageables. La première solution est d'imprimer le texte et de le numériser. Dans certains cas très rares, le fichier pourra également avoir été protégé contre l'impression. La seconde solution consiste à passer outre la protection numérique du document grâce à plusieurs logiciels de décryptage qui existent sur internet.

2. La deuxième cause possible à l'impossibilité de la copie des fichiers PDF au format texte réside dans le fait que les pages de texte ont été enregistrées sous la forme d'images. Ces images sont lisibles à l'œil, mais ne sont pas de l'information textuelle. Dans ce cas, il est conseillé d'utiliser un logiciel de reconnaissance optique des caractères. La plupart des logiciels récents peuvent ouvrir les fichiers PDF et en extraire l'information textuelle.

**Que faire dans le cas de documents audio-visuels et autres documents divers?**

Bien que SATO ne permette pas d'analyser directement les documents audio-visuels, il est possible de les codifier afin d'en effectuer l'analyse. La codification peut être une retranscription dans le cas d'un document audio. La codification peut également consister en une description détaillée dans le cas de photographies ou de documents visuels. Ainsi, on peut envisager l'utilisation de SATO pour l'analyse d'un corpus de descriptions de peintures ou d'œuvres musicales! Le problème de cette description reste cependant très complexe et renvoie à la compétence du musicologue et du critique d'arts.

**Quel nom de fichier devrait-on donner à ses documents?**

Comme expliqué dans la partie *Principes et méthodes* de ce Guide, un corpus est un ensemble raisonné de textes. Ce choix et cette organisation des textes devraient nous inciter à donner des noms judicieux aux fichiers qui contiennent les textes. L'homogénéisation des noms de fichiers a deux objectifs. Premièrement, elle permet d'identifier facilement le contenu des fichiers. C'est l'aspect documentaire. Secondement, elle facilite l'analyse du corpus dans SATO en permettant de sélectionner des documents dans le corpus selon des critères codés dans le nom des fichiers-documents. Ces critères permettent de désigner un certain nombre de variables externes caractérisant les conditions de production du texte. L'idée est de réserver des caractères dans le nom du document pour coder la provenance ou la nature du document. Il ne s'agit pas de tout coder, mais de se concentrer sur les éléments les plus pertinents pour l'analyse. Ces indications peuvent faire référence à la date, à la source, à l'auteur ou au numéro unique identifiant le texte.

Par exemple, si on travaille sur les discours des premiers ministres, on pourrait avoir des interventions provenant d'assemblées législatives (fédéral, provincial, municipal, etc.) et d'autres de conférences de presse. Ces informations pourraient être codées dans le nom des fichiers : *pm-fed-leg1998.txt pm-que-leg1999.txt*, *pm-fed-pr2000.txt*, etc. Ces noms feront référence aux discours des premiers ministres au parlement fédéral (*fed*) ou québécois (*que*) tels qu'ils ont été lus à l'assemblée législative (*leg*) ou en conférence de presse (*pr*).

Par exemple, si on travaille sur le racisme dans les médias, les noms de fichier pourraient avoir la forme suivante : *rac-pre1.txt*, *rac-pre2.txt*, *rac-dev1.txt*, *rac-jdm1.txt* et *rac-jdm2.txt*. Dans ce cas-ci, on s'est contenté de coder le thème de la sélection, le nom du journal et un numéro unique permettant de renvoyer à un répertoire bibliographique, par exemple. Ici *pre* fait référence au quotidien la Presse, *dev* au Devoir et *jdm* au Journal de Montréal. Dans ces exemples, nous avons choisi d'utiliser des tirets pour faciliter la lecture des noms de fichiers. Mais, on aurait pu faire plus court en les omettant dans la mesure où chacun des codes composant le nom débutent à une position fixe dans la chaîne de caractères.

Dans SATO, on pourra utiliser les noms des fichiers afin de sélectionner des documents. Par exemple, si, dans un projet sur la relation des jeunes à la politique au Québec, on a trente entrevues de jeunes qui se distinguent à partir de critères socio-économiques, on gagnera en facilité à identifier les entrevues en fonction de ces mêmes critères. Ainsi, par exemple les fichiers nommés *jeu-m-15-bas-parhorspol-03.txt*, *jeu-f-12-mtl-parenpol-18.txt* et *jeu-f-15-que-parpaspol-25.txt* pourraient s'interpréter de la façon suivante

- *jeu* identifie le thème des entrevues (*jeunes et politique*);
- *m* et *f* fait référence au sexe masculin ou féminin du répondant;

- les deux chiffres qui suivent donnent l'âge du répondant;
- les trois lettres suivantes indiquent d'où proviennent les jeunes, soit, du bas du fleuve (*bas*), de Montréal (*mtl*) ou de Québec (*que*);
- La cinquième section du nom fait référence à l'implication des parents en politique : *parenpol* et *parpaspol* pour parents impliqués en politique et parents n'ayant aucune implication en politique.

Ainsi nommés, il sera facile de sélectionner les documents selon les critères représentés. On pourra, par exemple, effectuer des analyses contrastives entre les garçons et filles de 15 ans et plus; entre les jeunes en provenance des centres urbains et ceux habitant en région; ou encore entre les filles ayant des parents impliqués en politique et ceux n'ayant aucun parent actif en politique.

Notons un petit détail si on veut inclure des dates dans les noms de fichiers. Dans ce cas, on aura avantage à utiliser un format normalisé du type année-mois-date (*2008-10-01*) plutôt qu'un format de type *01-oct-2008*. En effet, comme les ordinateurs trient les noms de fichier selon l'ordre alphabétique des caractères en commençant par la gauche, seul le premier format permettra de trier les dates dans l'ordre chronologique.

**Attention.** Tous ces exemples sont fournis à titre indicatif. Un fichier bien nommé permet, même des années après, de comprendre facilement la structure des données. Cependant, si la nomenclature est très complexe, il serait bien avisé de rédiger un fichier de type *lisez-moi.txt* afin de décrire la convention adoptée pour les noms de fichiers.

## 2.2. L'envoi des documents sur le serveur

## Le bureau de SATO

Le bureau de SATO est l'interface aux couleurs bleues-grises de SATO qui donne accès à divers outils d'édition et de création de corpus ainsi que de gestion des documents dans l'espace-usager sur le serveur.

Pour rentrer sur le bureau de SATO, il est nécessaire au préalable de s'inscrire. Une fois l'inscription complétée, l'usager pourra accéder à son espace privé. Ainsi, tous les documents et corpus qui se trouvent dans l'espace de l'usager sont privés et protégés par un mot de passe. Un deuxième mot de passe peut être créé pour donner accès à ses données en mode lecture aux autres utilisateurs qui connaissent le mot de passe de partage.

Les données analysées par SATO au travers de l'*architecture client-serveur* (c'est-à-dire en utilisant un navigateur web pour accéder à un serveur local ou distant) ne posent pas plus ni moins d'enjeux de confidentialité des données que l'utilisation du courriel en ligne par exemple (*Yahoo Mail*, *Hotmail* ou *Gmail*). Aussi, les serveurs SATO sont protégés physiquement et électroniquement contre les attaques les plus courantes.

Le bureau de SATO est divisé en deux parties principales : *Analyser* et *Outils (*champs 1 et 2 dans l'illustration suivante*)*. Les instructions de ce guide concernent avant tout la section *Outils*. La partie 1 du formulaire permet d'appeler les fonctions d'analyse de SATO sur un corpus existant.

Le bureau de SATO


## 2.3 La soumission d'un corpus à SATO

La soumission d'un corpus à SATO implique trois étapes :

1. L'envoi des textes dans l'espace privé de l'usager sur le serveur;
2. La documentation du corpus et de ses éventuelles propriétés;
3. La génération du corpus.

### 2.3.1 L'envoi des textes

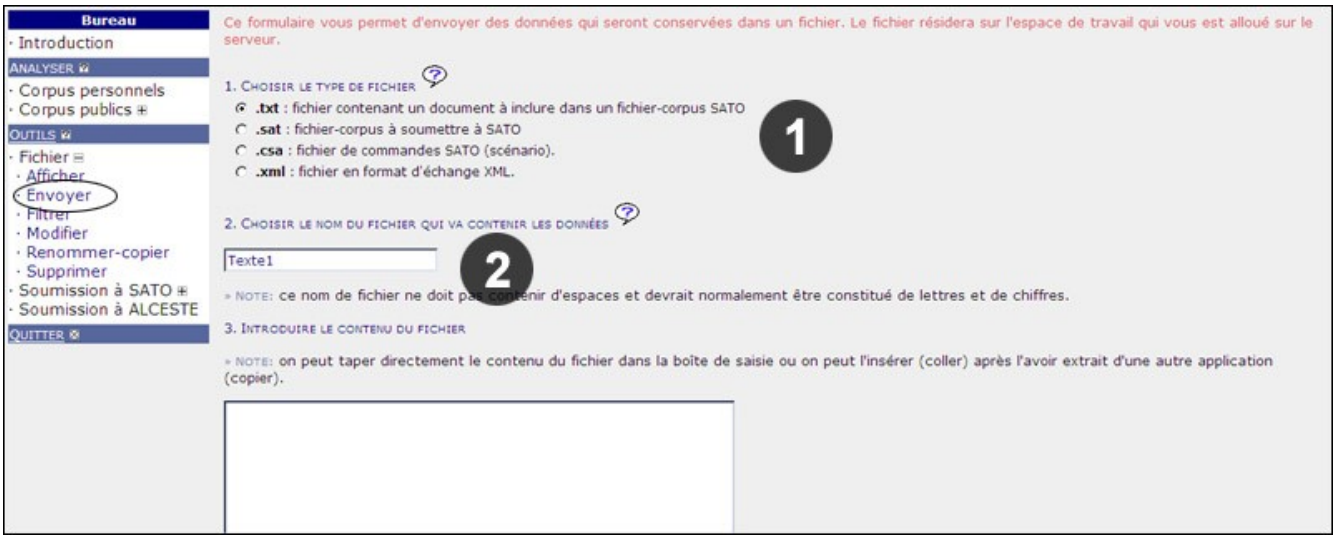
**Comment déposer des documents textuels dans son espace privé sur le serveur?**

Un corpus est généralement composé de plusieurs documents. Ces documents peuvent être rassemblés dans un seul fichier, chaque document étant introduit par la balise *page* présentée dans la partie *Principes et* méthodes de ce *Guide*. Dans la section *À quoi sert la propriété page?*, on indiquait aussi que la balise *page* disposait d'un l'opérateur @ permettant d'aller chercher le contenu du document dans un fichier déjà envoyé sur le serveur. En effet, si les documents à analyser ont une certaine envergure, ou s'ils sont susceptibles de faire partie de plusieurs corpus, il est préférable de les déposer séparément sur le serveur afin

qu'ils soient enregistrés dans des fichiers autonomes.

Actuellement, la meilleure manière pour un usager de déposer un document textuel dans son espace privé sur le serveur est d'utiliser le formulaire d'envoi de fichier. Ce formulaire est disponible sur le bureau sous la section *Outils*. On cliquera sur le lien *Fichier* et ensuite sur *Envoyer*.

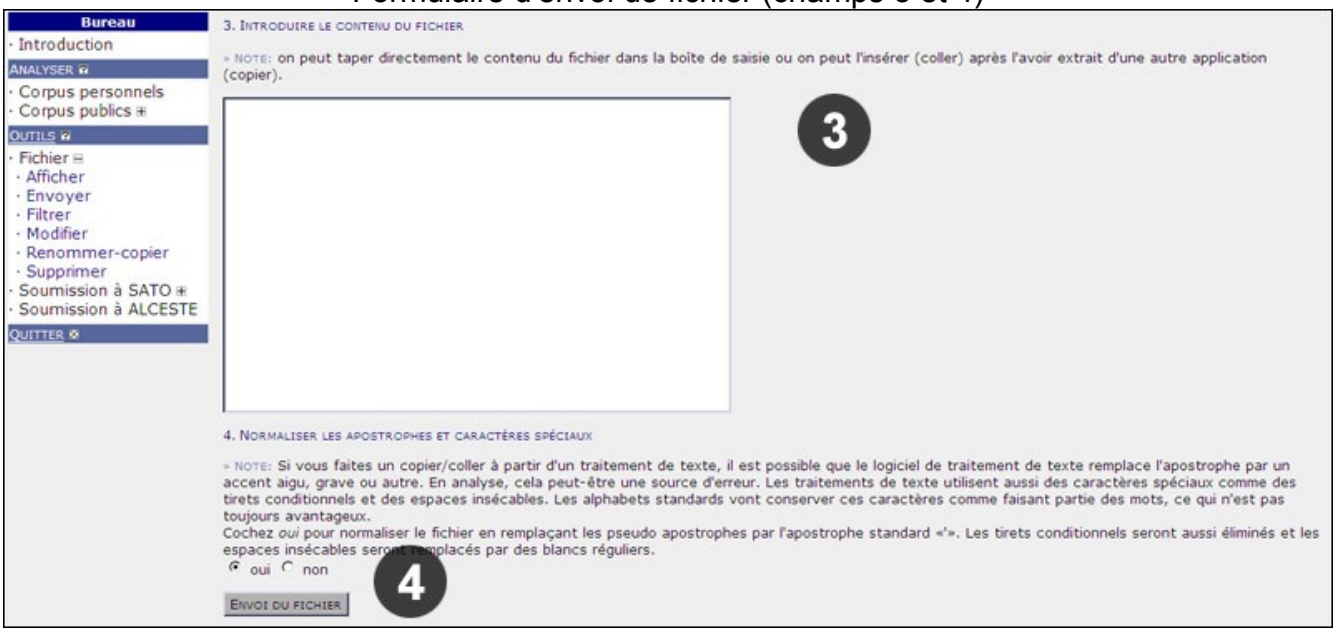
Formulaire d'envoi de fichier (champs 1 et 2)




Normalement, le nom de fichier (champ 1) devrait porter l'extension *.txt*. Dans le champ réservé au nom du fichier (champ 2), on mettra le nom sous lequel les données seront enregistrées sur le serveur.

**Attention.** Pour s'assurer que les noms de fichiers sont compatibles avec tous les systèmes d'exploitation, le nom des fichiers ne doit contenir que des caractères alphanumériques sans accents. On peut aussi utiliser les caractères – et _ à l'intérieur du nom. On notera, en particulier, qu'il ne faut pas mettre d'espaces dans les noms de fichiers.

Formulaire d'envoi de fichier (champs 3 et 4)




Dans le champ 3 (contenu du fichier), on collera le contenu du document que l'on aura sélectionné et copié dans le presse-papier depuis l'application qui affiche le texte sur son poste de travail (traitement ou éditeur de texte, navigateur Web, etc.).

Le champ 4 contient un bouton pour activer ou désactiver la normalisation des caractères. En effet, on retrouve souvent une variété de caractères pour représenter l'apostrophe. Pour normaliser les traitements, SATO propose une conversion vers l'apostrophe standard ('). Aussi, les traitements de texte utilisent souvent des espaces insécables pour attacher des ponctuations au mot qui le précèdent. Pour SATO, un tel espace insécable fera partie du mot, ce qui n'est probablement pas ce que l'on désire. L'option de normalisation transformera les espaces insécables en espaces standards. Il est suggéré de laisser l'option à *oui*, telle que sélectionnée dans le formulaire. Pour envoyer le document, on clique sur le bouton d'envoi.

On pourra répéter cette procédure pour chacun des documents textuels à traiter.

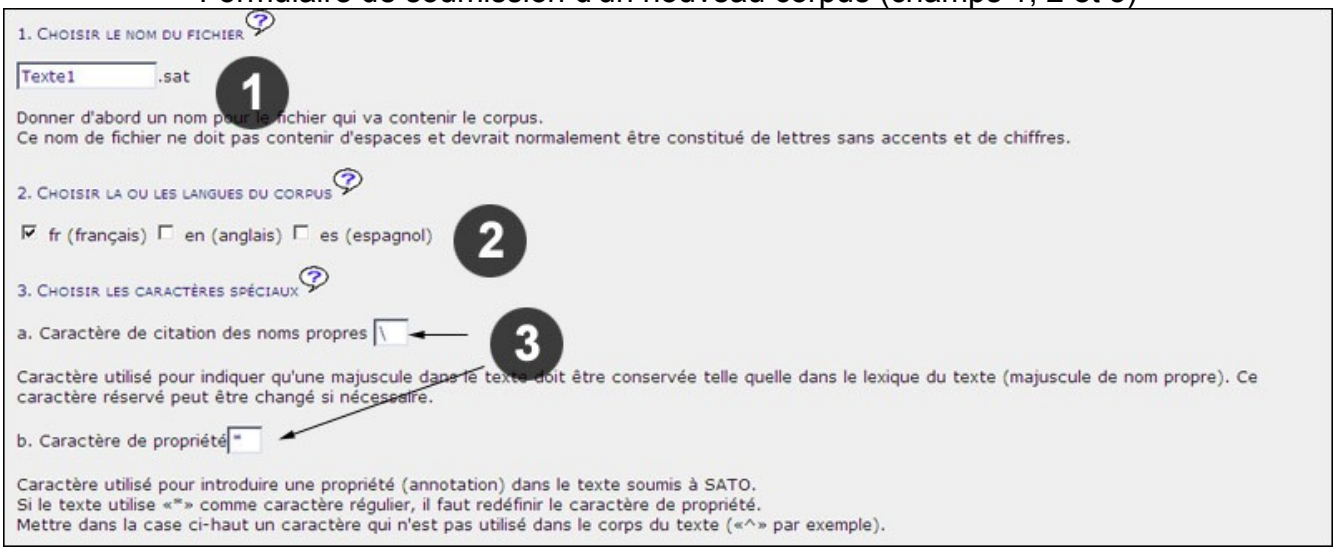
Pour les usagers devant manipuler un important volume de données, il est aussi possible de demander au Centre d'ATO l'ouverture d'un accès par FTP (*File Transfer Protocol*) à son espace privé sur le serveur. Ce protocole permet d'utiliser des logiciels de transfert de données (clients FTP) qui ressemblent aux explorateurs de fichiers de l'ordinateur personnel.

## 2.3.2. La documentation du corpus et de ses propriétés

Sous la section *Outils* du menu de gauche dans le bureau de SATO (l'interface bleue-grise), on trouve un lien intitulé *Soumission à SATO*. Un clic sur ce lien révèle trois nouveaux liens. Afin de soumettre un nouveau corpus, on utilisera le premier lien *Créer et soumettre un corpus*.

**Attention.** Sur la page intitulée *Créer et soumettre un corpus*, on trouvera une série de cases à remplir ou à cocher. Afin d'obtenir de l'aide à propos de chacune des cases à remplir, on cliquera sur l'icône en forme de bulle avec un point d'interrogation.

Formulaire de soumission d'un nouveau corpus (champs 1, 2 et 3)

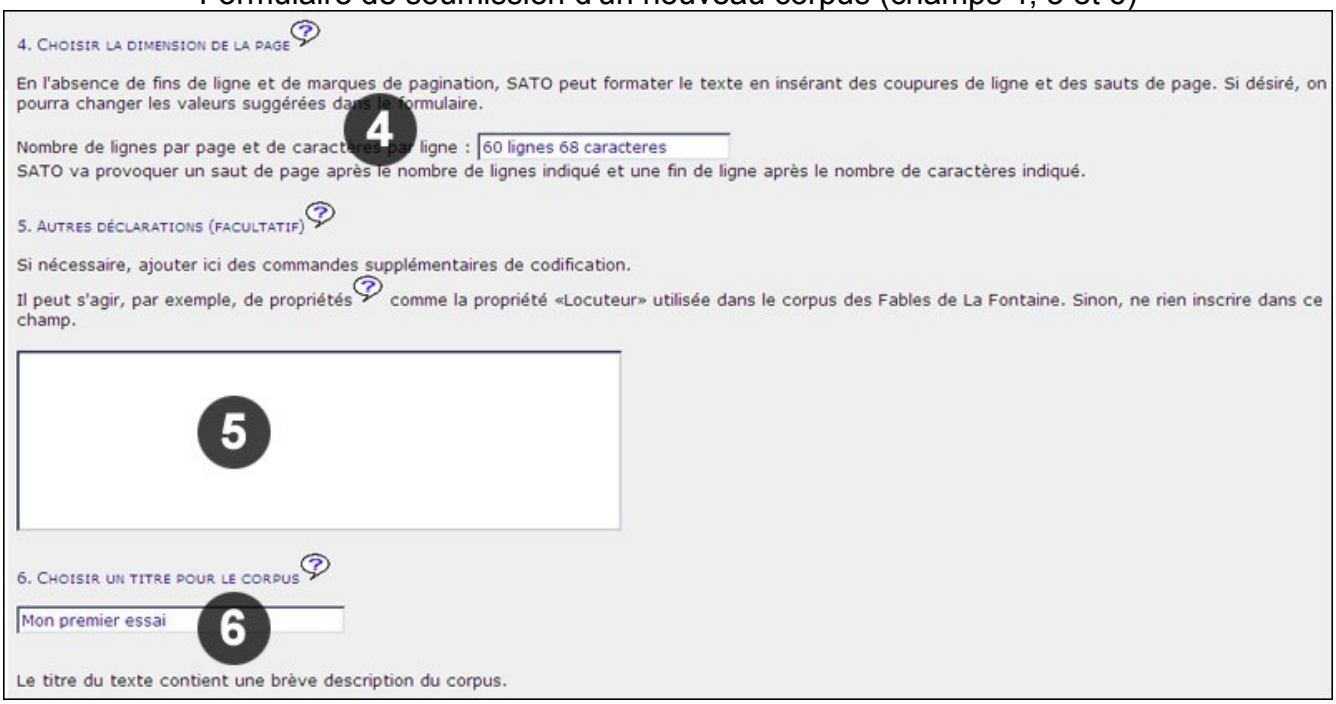



Le premier champ du formulaire contient le nom du fichier-corpus rassemblant les divers documents à analyser. L'extension *.sat* sera ajoutée automatiquement. Comme partout dans SATO, les noms de fichiers ne doivent contenir que des caractères alphanumériques sans accents.

Le deuxième champ du formulaire permet de choisir un ou plusieurs alphabets standards correspondant aux langues utilisées dans le corpus. Le Manuel SATO explique comment faire pour définir de nouveaux alphabets (à insérer dans le champ 5).

Le troisième champ du formulaire permet de définir les caractères spéciaux qui seront interprétés de façon particulière lors de la lecture des textes constituants le corpus. Le caractère de citation des noms propres (\ si on ne le change pas) indiquera à SATO que le caractère suivant devra être pris tel sans conversion des majuscules. Aussi le caractère de propriété (* s'il n'est pas redéfini) permet d'introduire une annotation sous forme de propriété SATO.

Formulaire de soumission d'un nouveau corpus (champs 4, 5 et 6)




Le champ 4 permet de définir les paramètres se segmentation automatique en lignes et en pages. Cette fonction est surtout utile pour des textes continus sans fins de lignes. Si les textes du corpus sont déjà paginés avec des retours à chaque fin de ligne, on doit changer les paramètres, par exemple des pages de 1000 lignes et les lignes de 1000 caractères, afin d'éviter que la pagination automatique n'entre en conflit avec la pagination existante.

Le champ 5 contiendra les définitions des alphabets supplémentaires s'il y a lieu. Mais, surtout, on indiquera dans *Autres déclarations* quelles propriétés, outre les propriétés prédéfinies, ont été utilisées pour le balisage du corpus.

On renseignera donc les propriétés ainsi :

> Propriété [nom de la propriété] [type de propriété (*entière*, *libre* ou *symbolique*] [portée de la propriété (*texte* ou *lexique*)] [toutes les valeurs attribuées]

1. Premièrement, on introduit à SATO le fait que la ligne va porter sur une propriété en entrant le mot *propriété*.

2. Ensuite, on indique à SATO de quelle propriété ou balise il s'agit. On indique ici le nom choisi. Dans les exemples précédents, nous avions choisi *partie*, *classification*, *sens*, *thématique* ou encore *locuteur*. Ce sont les noms des propriétés.

3. En troisième lieu, on indique à SATO le type de la propriété. Est-ce que les valeurs permises pour la propriété sont des valeurs libres, entières ou symboliques?

4. En quatrième lieu, on spécifie la portée de la propriété. Est-ce que la propriété s'applique au texte ou au lexique?

5. Pour finir, dans le cas des propriétés symboliques, on fait la liste de toutes les valeurs possibles de la propriété, sauf *nil* qui est préfini et signifie *aucune valeur*. On sépare toutes les valeurs symboliques par un espace sans utiliser de virgule ni quelqu'autre signe que ce soit. L'ordre des valeurs n'a pas d'importance. En fait, la liste des valeurs de la propriété est facultative. Si la liste est fournie, SATO pourra en valider l'application et signaler une erreur d'orthographe, par exemple. Si aucun symbole n'est donné, SATO les apprendra au fur et à mesure. Si on fait une faute d'orthographe dans le nom de la valeur, elle sera aussi apprise. Il faudra donc valider la liste des valeurs en première analyse.

Voici comment déclarer les propriétés que nous avons utilisées dans des exemples précédents.

**Exemple 1**

Utilisation dans le corpus.

***thématique=biographie**
***{**bande_1 400 s.**}**
***locuteur=I1** Qu'est-ce qui dans ton enfance t'a porté à t'intéresser à l'écologie?

***locuteur=T1** L'entrevue a été interrompu par la sonnerie du téléphone. Le répondant décide de ne pas répondre.

***locuteur=Rf22** Eh bien… J'avais… 8 ou 9 ans je crois. Je me rappelle qu'un ami m'avait invité chez eux. Ses parents faisaient partie d'une organisation écologique. Lorsque je suis arrivé chez eux, les parents discutaient d'une action qu'ils allaient mener contre une usine qui déversait des produits chimiques dans une rivière proche. […]

***thématique=opinion**
***{**bande_2 320 s.**}**
***locuteur=I1** Qu'est-ce que tu penses de ceux qui disent qu'on ne devrait pas impliquer les enfants en bas âge en politique?

***locuteur=Rf22** Je ne suis pas d'accord avec eux _soupir. _silence Je trouve que de toutes manières les enfants sont entourés de situations dans lesquelles il est important d'avoir une idée de ce que doit être la société. Et que de les impliquer le plus tôt possible est une des meilleures manières de les préparer à la vie en société.

Déclaration lors de la documentation des propriétés

```
Propriété thématique symbolique pour texte biographie opinion

Propriété locuteur symbolique pour texte I1 T1 Rf22
```

**Exemple 2**

Utilisation dans le corpus.

```
Quand je vais à la pêche, je n'oublie jamais d'emporter mes
vers*sens="animal". Lorsque je les accroche à l'hameçon, je tâche de
les installer vers*sens="direction" le centre.
```

Déclaration lors de la documentation des propriétés.

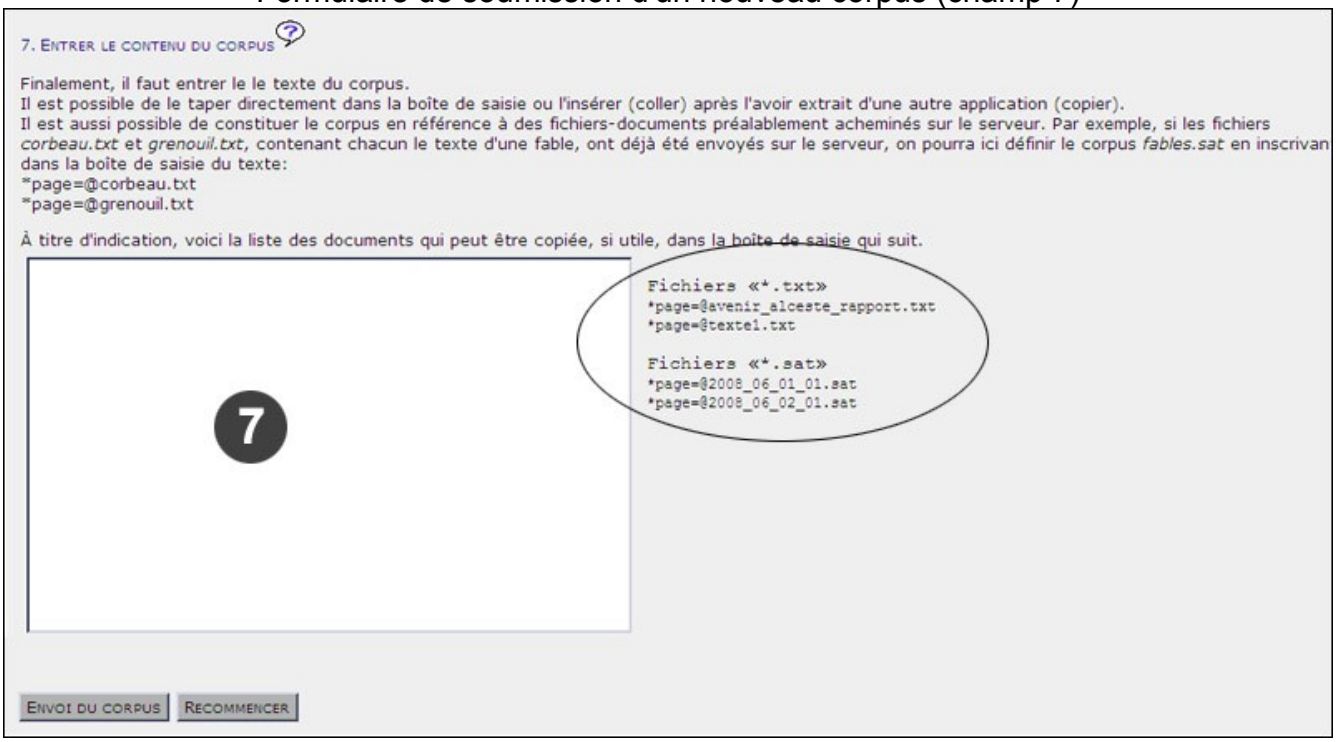
```
Propriété sens symbolique pour lexique animal direction
```

On notera qu'il n'est pas nécessaire d'ajouter *nil* aux déclarations de propriétés symboliques puisque c'est la valeur par défaut de toutes les propriétés symboliques. Dans le cas où on attribue plus d'une propriété au corpus, chaque propriété déclarée occupe sa propre ligne.

**Attention.** La déclaration d'une propriété doit se faire de manière continue. Elle peut s'effectuer sur plusieurs lignes mais elle ne peut pas contenir de retour de chariot. Si, pour des raisons particulières, on désire travailler sur plusieurs lignes, il faudra insérer deux étoiles « ** » à la fin de la ligne avant le retour de chariot. SATO comprendra ainsi que la commande de déclaration n'est pas terminée et qu'il faut considérer la ligne suivante comme faisant partie de la ligne précédente.

Le champ 6 permet de donner un titre significatif au corpus.

Formulaire de soumission d'un nouveau corpus (champ 7)




Le formulaire d'envoi se termine par le contenu textuel du corpus. On peut copier directement le contenu du texte qu'on aura collé à partir d'une application ouverte sur son ordinateur. Mais, on peut aussi insérer un document déjà envoyé sur le serveur. On trouvera la liste des documents envoyés à droite de la boîte d'entrée du champ 7. Dans ce cas, on copie dans la boîte de saisie les noms des fichiers au complet (c'est-à-dire tels qu'ils sont précédés de *page=@). Normalement, les documents à insérer devraient porter l'extension *.txt*. Mais, SATO liste aussi les documents portant l'extension *.sat*.

Une fois le descriptif du corpus envoyé sur le serveur, on pourra le soumettre à SATO pour qu'il le *génère*.

## 2.3.3. La génération du corpus

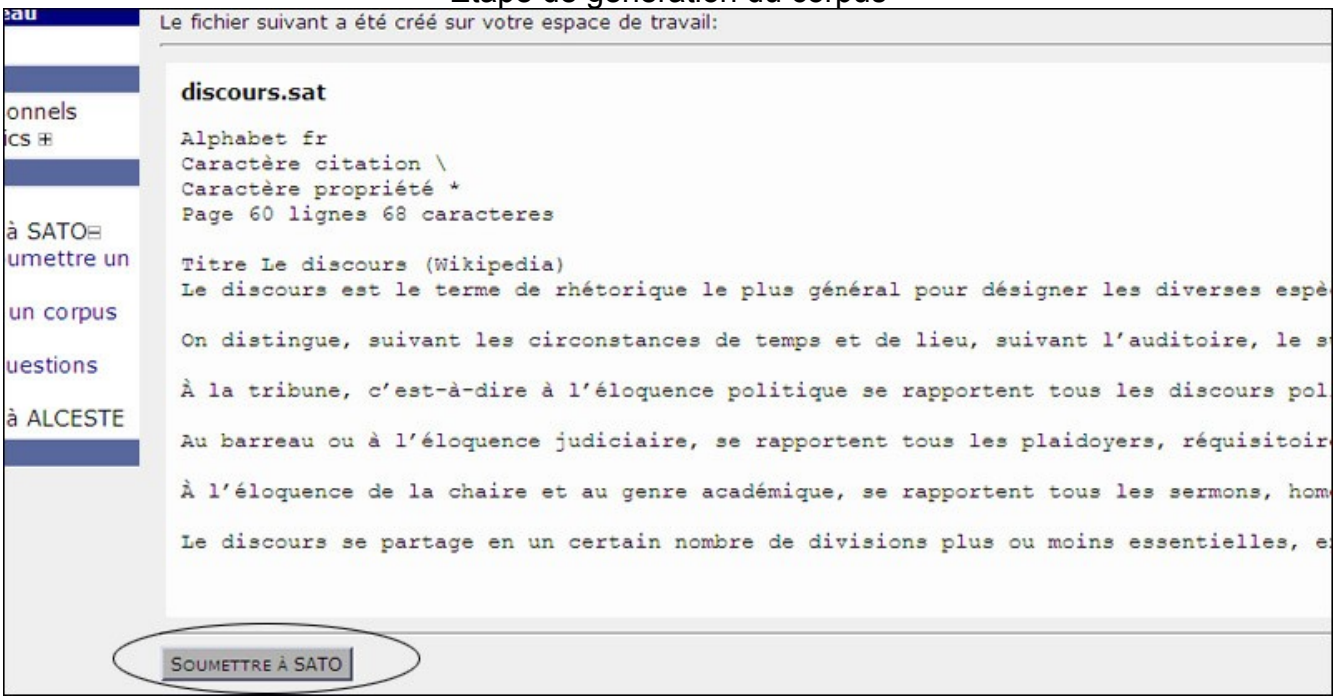
### Qu'est-ce que la *SATO génération* d'un corpus?

**Définition.** La génération d'un corpus par SATO est la transformation de celui-ci en un format que le logiciel peut interpréter. En particulier, cette transformation permet de générer le lexique des mots du texte et d'affecter des valeurs de propriétés aux formes lexicales ou à leurs occurrences dans le corpus.

L'étape de génération d'un corpus peut être très rapide pour les petits corpus, mais elle peut prendre jusqu'à quelques minutes pour les très gros corpus. Durant cette étape, SATO vérifie la syntaxe des méta-données. Il construit un répertoire des mots du corpus et il attribue à tous les mots du corpus un numéro unique.

La génération est l'étape qui suit le clic sur le bouton *Envoi du corpus* en bas du formulaire *Créer et soumettre un corpus* de la section *Outils* du bureau de SATO.

Étape de génération du corpus




**Attention.** Il arrive souvent que les débutants dans SATO génèrent, par erreur, leur corpus à chaque fois qu'ils désirent y accéder. À moins qu'on doive rajouter ou enlever du texte, ou qu'on doive corriger des fautes dans le corpus, on effectue **une seule** génération de corpus. Sinon, le travail d'enrichissement du corpus réalisé en analyse sera écrasé et perdu! Donc, à moins de devoir modifier un corpus déjà généré, on y accèdera à partir de la section *Analyser* du bureau de SATO (interface bleue-grise).

Formulaire de sélection d'un corpus à analyser


Les corpus offerts pour l'analyse sont les corpus déjà soumis avec succès à SATO. Il peut

s'agir de corpus présents dans son espace personnel ou dans un espace partagé si on a choisi cette option lors de l'ouverture de la session. On peut aussi accéder à un corpus public si on clique sur l'onglet *Corpus publics* du formulaire.

**Que faire s'il on rencontre un problème durant la génération d'un corpus?**

Il existe une Foire aux questions (FAQ) consacrée à la soumission de corpus. Elle est disponible dans la section *Outils* du bureau de SATO sous le lien *Soumission à SATO*.

En général, ce sont les problèmes liés aux propriétés qui empêchent SATO de générer correctement le corpus. Or, tant que le corpus n'a pas été généré sans erreurs, il ne peut pas être analysé. En cas d'erreur, un message est alors affiché. Il s'agit d'en prendre note et de corriger le corpus en conséquence. L'onglet *Modifier* sous la rubrique *Fichier* dans le menu de gauche du bureau permet d'accéder à ses données pour les corriger. Dans le *patron de sélection des fichiers* du premier formulaire de modification, on choisira **.txt* pour accéder à ses fichiers-documents et **.sat* pour accéder aux définitions des corpus avec ses déclarations de propriété. Ensuite, il s'agira de reprendre l'étape de soumission du corpus par l'onglet *Soumettre un corpus déjà existant* sous la rubrique *Soumission à SATO* du bureau.·

## 2.4 Identification et correction des problèmes les plus courants

Avant de commencer l'analyse approfondie du corpus, il est conseillé de procéder à une analyse exploratoire afin de vérifier que toutes les balises ont été employées correctement. Cette section du *Guide* propose quelques trucs pour explorer le corpus afin de détecter les problèmes les plus fréquents. Elle explique également comment corriger ces problèmes.

Voici les quelques manipulations que nous conseillons d'effectuer afin de trouver les problèmes les plus courants d'un corpus nouvellement soumis.

**Description du texte et du lexique**

Tous les documents du corpus ont-ils été reconnus par SATO? Est-ce que toutes les pages de chacun des documents ont bien été dépistées? Il peut arriver que, lors de la préparation du corpus, certains documents aient été oubliés ou que certaines balises de type **page* ne soient pas correctement orthographiées.

Est-ce que le nombre de mots détectés correspond à l'ordre de grandeur du corpus? Un commentaire mal fermé, par exemple, pourrait avoir pour effet de faire disparaître des milliers de mots! Heureusement, SATO émet un avertissement lorsqu'il rencontre une balise d'ouverture de commentaire à l'intérieur d'un commentaire (*{ ).

Est-ce que le lexique du corpus est représentatif de la répartition des mots selon les langues utilisées?

Deux commandes générales de SATO permettent de répondre à ces questions. Il s'agit de *Texte décrire* et *Lexique décrire*. Voici un exemple tiré du corpus public sur le discours

constitutionnel canadien (DCC).

TEXTE DECRIRE

```
Projet Discours constitutionnel canadien (1941-1987)
--texte généré le 20-9-2006 à 9:58:48 heure par SATO version 4.21

Alphabet fr a b c d e f g h i j k l m n o ¤ p q r s t u v w x y z 0 ,
0 .0 ¼ ½ ¾ 1 ,1 .1 ¹ 2 ,2 .2 ² 3 ,3 .3 ³ 4 ,4 .4 5 ,5 .5 6 ,6 .6 7 ,7 .
7 8 ,8 .8 9 ,9 .9 *accent '_ _ ~ ^ ° - *séparateur , : ; . ? ¿ ! ... <
> ( ) [ ] { } « » % $ £ ¢ ¥ # " @ & + = / \ | * ÷ ± ®  ¦ *terminal ' ª
º


INFORMATION Ce corpus a été préparé par le Groupe de recherche en
analyse du discours politique (GRADiP) de l'UQAM (Université du Québec
à Montréal). Nous rendons disponible à la consultation et à l'analyse
cet ensemble textuel, le discours constitutionnel canadien qui a fait
par ailleurs l'objet de la publication du livre « L'identité fragmentée
» écrit par Gilles Bourque et Jules Duchastel avec la collaboration de
Victor Armony récipiendaire du prix Richard-Arès en 1996. Vous
trouverez plus d'information concernant le corpus, sa constitution et
les catégories utilisées sur le site suivant: http://ato.chaire-mcd.ca/

Nombre de mots dans le texte: 368707
Nombre de lignes: 45888

Document dcc-1941: page(s) 1-42
Document dcc-1945: page(s) 1-145
Document dcc-1950: page(s) 1-73
Document dcc-1964: page(s) 1-26
Document dcc-1968: page(s) 1-84
Document dcc-1969: page(s) 1-98
Document dcc-1971: page(s) 1-42
Document dcc-1978: page(s) 1-17
Document dcc-1980: page(s) 1-48
Document dcc-1981: page(s) 1-28
Document dcc-1983: page(s) 1-40
Document dcc-1984: page(s) 1-53
Document dcc-1985: page(s) 1-30
Document dcc-1987: page(s) 1-47

nombre total de documents: 14
nombre total de pages: 773
```

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Texte* dans le menu de gauche de l'interface avancée;
2) Choisir *Décrire.*

La description du texte permet d'avoir une idée générale de la structure du corpus. Voici quelques questions qu'on pourrait se poser en lisant la description du texte :

- Est-ce que tous les documents du corpus ont été reconnus par SATO?
- Est-ce que toutes les pages de chacun des documents ont bien été dépistées?
- Est-ce que le nombre de mots détectés correspond à l'ordre de grandeur du corpus?

Voici maintenant la description du lexique tiré du corpus DCC.

LEXIQUE DECRIRE

```
Alphabet fr a b c d e f g h i j k l m n o ¤ p q r s t u v w x y z 0 ,
0 .0 ¼ ½ ¾ 1 ,1 .1 ¹ 2 ,2 .2 ² 3 ,3 .3 ³ 4 ,4 .4 5 ,5 .5 6 ,6 .6 7 ,7 .
7 8 ,8 .8 9 ,9 .9 *accent '_ _ ~ ^ ° - *séparateur , : ; . ? ¿ ! ... <
> ( ) [ ] { } « » % $ £ ¢ ¥ # " @ & + = / \ | * ÷ ± ®  ¦ *terminal ' ª
º


INFORMATION Ce corpus a été préparé par le Groupe de recherche en
analyse du discours politique (GRADiP) de l'UQAM (Université du Québec
à Montréal). Nous rendons disponible à la consultation et à l'analyse
cet ensemble textuel, le discours constitutionnel canadien qui a fait
par ailleurs l'objet de la publication du livre « L'identité fragmentée
» écrit par Gilles Bourque et Jules Duchastel avec la collaboration de
Victor Armony récipiendaire du prix Richard-Arès en 1996. Vous
trouverez plus d'information concernant le corpus, sa constitution et
les catégories utilisées sur le site suivant: http://ato.chaire-mcd.ca/

Nombre de formes lexicales: 17682
```

Pour obtenir ce résultat à partir de l'interface Web, on doit :
1) Cliquer sur le lien *Lexique* dans le menu de gauche de l'interface avancée;
2) Choisir *Décrire.*

La description du lexique permet d'avoir une idée générale de l'utilisation des divers alphabets dans un corpus multilingue. Dans l'exemple du corpus DCC, qui est uniquement en français, la seule chose nouvelle révélée par la commande est le nombre de formes lexicales dans le lexique du corpus.

**Description des propriétés**

Pour vérifier le balisage du corpus avec diverses propriétés, on peut d'abord vérifier les définitions des propriétés par la commande *Propriété afficher*.

PROPRIÉTÉ AFFICHER

```
propriété NoLex entière pour lexique

propriété NoOcc entière pour texte
```

```
propriété Alphabet symbolique pour lexique fr

propriété Édition symbolique pour texte collé lié maj par cap tab 2 3 4
5 6 7 8 np fdl

propriété Fréqtot entière pour lexique

propriété Longueur entière pour lexique

propriété Page référentielle pour texte

propriété Commentaire libre pour texte

propriété locuteur symbolique pour texte alb-gett alb-loug alb-mann
alb-stro aut-bruy aut-cian cb-benne cb-johns cb-pattu cb-zalm féd-ilsl
féd-mack féd-mulr féd-pear féd-stla féd-trud inu-amag inu-cian inu-nung
ipe-camp ipe-ghiz ipe-jone ipe-lee ipe-macl ipe-shaw man-camp man-gars
man-lyon man-pawl man-robl man-schr man-weir nb-hatfi nb-mcnai nb-robic
ne-bucha ne-macdo ne-macmi ne-morri ne-regan ne-smith ne-stanf ont-davi
ont-fros ont-mill ont-pete ont-roba qué-bert qué-bour qué-dupl qué-godb
qué-inco qué-john qué-lesa qué-léve sas-blak sas-devi sas-doug sas-heal
sas-steu sas-that tn-peckf tn-small

propriété socio symbolique pour lexique ec0 ec1 ec12 ec13 ec14 ec15
ec19 ec19a ec2 ec20 ec3 ec4 ec5 ec6 ec7 et0 et1 et1a et2 et2a1 et2a2
et2b1 et2b2 et2b3 et2b4 et2c et2c1 et2c2 et2c3 et2c4 et2c5 et2d et3 et4
et5 et6 et7 et7a in0 in1 in10 in2 in3 in4 in5 in6 in7 in8 in8a in8b in9
rien us0 us1 us10 us11 us12 us13 us1a us1b us1c us2 us2a1 us2a2 us2a3
us2a4 us2a5 us2b1 us2b2 us2c1 us2c2 us2c3 us2c4 us2c5 us2c6 us2c7 us3
us4 us5 us6 us7 us7a1 us7a2 us7a3 us7b1 us7b2 us7b3 us7b4 us7b5 us7b6
us7b7 us7b8 us7b9 us8 us8a us8b us8c us9 uv1 uv10 uv11 uv11a uv12 uv13
uv14 uv15 uv16 uv17a uv17b uv17c uv18 uv19 uv1a uv2 uv20 uv21 uv22 uv23
uv24 uv25 uv26 uv27 uv27a uv27b uv28 uv3 uv31 uv32 uv33 uv34 uv4 uv5
uv6 uv7 uv8 uv9
```

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Propriété* dans le menu de gauche de l'interface avancée;
2) Choisir *Afficher;*
3) Laisser le nom de la propriété en blanc pour obtenir la définition de toutes les propriétés.

Dans l'exemple du corpus DCC, outre les propriétés prédéfinies de SATO, la commande nous révèle les définitions des propriétés *locuteur* et *socio*. La propriété *locuteur* faisait partie de la transcription des discours alors que la propriété *socio* a été ajoutée en cours d'analyse à titre de catégorisation socio-sémantique des énoncés.

Il est aussi possible d'obtenir une statistique descriptive qui donnera un indice de l'utilisation d'une propriété. La manière rapide de vérifier la consistance d'une propriété est d'utiliser la commande *Propriété décrire*. Voici, pour le *Discours constitutionnel canadien*, la description de la propriété *locuteur*

Description de la propriété locuteur

filtre: $
Nombre de lexèmes sélectionnés: 17682/17682 (100.00 %)
Nombre d'occurrences sélectionnés: 368707

| Occ. | %Occ. | Seg. | %Seg. | locuteur |
|---|---|---|---|---|
| 35702 | 9.68% | 4 | 2.48% | "man-gars" |
| 20863 | 5.66% | 12 | 7.45% | "cb-benne" |
| 19909 | 5.40% | 4 | 2.48% | "féd-mack" |
| 16643 | 4.51% | 8 | 4.97% | "féd-trud" |
| 16394 | 4.45% | 4 | 2.48% | "ipe-jone" |
| 15060 | 4.08% | 6 | 3.73% | "qué-léve" |
| 12671 | 3.44% | 5 | 3.11% | "tn-peckf" |
| 12097 | 3.28% | 7 | 4.35% | "tn-small" |
| 10710 | 2.90% | 4 | 2.48% | "ont-roba" |
| 10620 | 2.88% | 5 | 3.11% | "qué-dupl" |
| 8825 | 2.39% | 3 | 1.86% | "nb-robic" |
| 8678 | 2.35% | 3 | 1.86% | "féd-stla" |
| 7806 | 2.12% | 4 | 2.48% | "ont-davi" |
| 7647 | 2.07% | 2 | 1.24% | "alb-stro" |
| 7393 | 2.01% | 3 | 1.86% | "alb-mann" |
| 6725 | 1.82% | 4 | 2.48% | "man-weir" |
| 6699 | 1.82% | 3 | 1.86% | "aut-bruy" |
| 6404 | 1.74% | 3 | 1.86% | "sas-devi" |
| 6260 | 1.70% | 4 | 2.48% | "alb-loug" |
| 6046 | 1.64% | 4 | 2.48% | "féd-pear" |
| 5924 | 1.61% | 2 | 1.24% | "ne-smith" |
| 5893 | 1.60% | 2 | 1.24% | "féd-mulr" |
| 5653 | 1.53% | 1 | 0.62% | "sas-doug" |
| 5518 | 1.50% | 2 | 1.24% | "man-lyon" |
| 5306 | 1.44% | 5 | 3.11% | "ne-bucha" |
| 5112 | 1.39% | 2 | 1.24% | "sas-blak" |
| 5015 | 1.36% | 4 | 2.48% | "man-pawl" |
| 4771 | 1.29% | 4 | 2.48% | "nb-hatfi" |
| 4526 | 1.23% | 1 | 0.62% | "sas-steu" |
| 4028 | 1.09% | 2 | 1.24% | "qué-john" |
| 3879 | 1.05% | 1 | 0.62% | "qué-bert" |
| 3744 | 1.02% | 3 | 1.86% | "ipe-camp" |
| 3607 | 0.98% | 2 | 1.24% | "cb-pattu" |
| 3550 | 0.96% | 1 | 0.62% | "féd-ilsl" |
| 3543 | 0.96% | 2 | 1.24% | "ne-macdo" |
| 3353 | 0.91% | 1 | 0.62% | "nb-mcnai" |
| 3289 | 0.89% | 2 | 1.24% | "sas-that" |
| 3212 | 0.87% | 2 | 1.24% | "ipe-lee" |
| 3162 | 0.86% | 2 | 1.24% | "ipe-macl" |
| 3078 | 0.83% | 1 | 0.62% | "man-schr" |
| 2972 | 0.81% | 2 | 1.24% | "ont-fros" |
| 2874 | 0.78% | 1 | 0.62% | "inu-cian" |
| 2446 | 0.66% | 1 | 0.62% | "alb-gett" |
| 2349 | 0.64% | 1 | 0.62% | "cb-zalm" |
| 2241 | 0.61% | 1 | 0.62% | "inu-nung" |
| 2052 | 0.56% | 1 | 0.62% | "ne-regan" |

| | | | | |
|---|---|---|---|---|
| 2031 | 0.55% | 1 | 0.62% | "ont-pete" |
| 1949 | 0.53% | 1 | 0.62% | "qué-bour" |
| 1926 | 0.52% | 1 | 0.62% | "sas-heal" |
| 1900 | 0.52% | 1 | 0.62% | "aut-cian" |
| 1855 | 0.50% | 1 | 0.62% | "man-camp" |
| 1847 | 0.50% | 1 | 0.62% | "ipe-ghiz" |
| 1778 | 0.48% | 2 | 1.24% | "ne-macmi" |
| 1745 | 0.47% | 2 | 1.24% | "inu-amag" |
| 1590 | 0.43% | 1 | 0.62% | "ipe-shaw" |
| 1234 | 0.33% | 1 | 0.62% | "qué-lesa" |
| 1127 | 0.31% | 2 | 1.24% | "qué-godb" |
| 1080 | 0.29% | 1 | 0.62% | "ne-morri" |
| 1053 | 0.29% | 1 | 0.62% | "ont-mill" |
| 941 | 0.26% | 1 | 0.62% | "ne-stanf" |
| 920 | 0.25% | 1 | 0.62% | "man-robl" |
| 838 | 0.23% | 1 | 0.62% | "cb-johns" |
| 644 | 0.17% | 1 | 0.62% | "qué-inco" |

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Propriété* dans le menu de gauche de l'interface avancée;
2) Choisir *Décrire;*
3) Ensuite, indiquer la propriété à décrire.

Pour chaque locuteur, on a un symbole qui décrit la province et le premier ministre qui a pris la parole. On a le nombre de mots prononcés par chacun et le nombre de segments, c'est-à-dire d'interventions continues. Ces chiffres sont aussi exprimés en pourcentage. La lecture de ce tableau permet de voir rapidement si le balisage du corpus, en termes de locuteurs dans notre exemple, couvre adéquatement l'ensemble des données.

**Exploration du texte**

L'affichage du texte intégral permettrait d'en vérifier l'intégrité. Mais lire tout le corpus peut être une opération fastidieuse. Il suffit souvent de parcourir quelques pages du corpus pour vérifier que l'opération de soumission à SATO s'est bien déroulée.

Pour faire afficher le texte à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Texte* dans le menu de gauche de l'interface avancée;
2) Choisir *Afficher;*
3) Ensuite, entrer *$* dans le champ du filtre.

Comme le texte s'affiche par partie avec contrôle d'écran et qu'il faut cliquer sur l'icône de déroulement pour passer à la suite de l'affichage, on pourra interrompre l'affichage et passer à une autre commande lorsque l'inspection visuelle du texte aura été jugée satisfaisante.

Si on a déjà détecté des anomalies dans la génération du corpus en examinant les résultat des commandes de description, on peut modifier le filtre dans l'affichage afin de se positionner sur un document ou une page en particulier. Par exemple, *$*page=doc1/5* afficherait la cinquième page du document *doc1*.

Il est aussi possible de colorer le texte en fonction des diverses valeurs d'une propriété symbolique afin de vérifier visuellement le balisage. Pour les détails, on pourra consulter le Manuel de référence SATO (commande *Poste écran caractériser couleur*). Ainsi, coloré, le texte affiché pourra, d'un coup d'oeil, révéler une anomalie de balisage.

**Exploration du lexique**

L'exploration du lexique du corpus permet d'identifier assez facilement les mots à problème, en particulier les mots mal orthographiés apparus lors de la numérisation de textes ou de la transcription d'entrevues. SATO propose diverses manières rapides d'explorer le lexique du corpus.

**Attention.** Ces trucs sont des solutions partielles et ne dispensent pas totalement d'un examen plus approfondi du lexique. Si on veut procéder à une vérification méticuleuse, on peut afficher tout le lexique. Comme on peut cliquer sur les entrées du lexique et faire afficher les contextes courts (*kwic*), cette lecture du lexique est souvent une première façon de se familiariser avec ses données.

Problème 1 : Les mots collés

Il peut arriver que, lors de l'encodage du corpus, certains mots se soient collés, faute d'espaces séparateurs.

COMMENT DÉTECTER CE PROBLÈME?

Une manière simple de détecter la présence de nombreux mots collés est d'afficher le lexique en le triant par ordre décroissant de longueur de mots. En effet, dans les langues non agglutinantes comme le français, l'anglais ou l'espagnol, les mots sont relativement courts. Les mots collés par inadvertance sont presque toujours parmi les mots les plus longs du corpus. Voici la commande à effectuer pour afficher les mots du lexique par ordre décroissant de longueur :

LEXIQUE AFFICHER $ TRI Longueur

Fréqtot

```
1          certaineméthode
3          circonstances
3          dissertations
1          divisionsplus
3          proclamations
3          réquisitoires
3          compositions
3          confirmation
3          essentielles
```

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Lexique* dans le menu de gauche de l'interface avancée;
2) Choisir *Afficher;*
3) Ensuite, entrer *$* dans le champ du filtre et choisir *longueur* comme propriété de tri;

Dans cet exemple, on dépiste deux erreurs liées à des mots collés : *certaineméthode* (certaine méthode) et *divisionsplus* (divisions plus).

Problème 2 : Des chiffres au lieu de lettres

Un deuxième problème que peut révéler l'examen du lexique, c'est la confusion entre chiffres et lettres dans certains mots. Certains mots contiennent quelques chiffres alors que ceux-ci devraient être des lettres. Par exemple, le chiffre « 1 » peut avoir été utilisé à la place de la lettre « l » ou encore, le chiffre « 0 » est parfois présent à la place de la lettre « O ». De plus, il se peut parfois que lors de la numérisation les appels de notes de bas de page se soient collés au mot qui le précède. Ainsi, dans l'exemple de la phrase suivante, le chiffre « 12 » est collé au mot « global » : « Actuellement, peu nombreux sont les scientifiques qui nient le réchauffement global[12]. » Ce problème, difficile à détecter à l'œil nu, est surtout fréquent dans le cas de textes numérisés.

COMMENT DÉTECTER CE PROBLÈME?

Une manière simple de détecter la présence de chiffres à la place de lettres est d'afficher le lexique de tous les lexèmes contenant des chiffres. La séquence suivante les affichera :

LEXIQUE AFFICHER $(1,2,3,4,5,6,7,8,9,0)$ TRI Alphabet

| Fréqtot | |
|---|---|
| 1 | 0raisons |
| 1 | d1scours |

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Lexique* dans le menu de gauche de l'interface avancée;
2) Choisir *Afficher;*
3) Ensuite, entrer *$(1,2,3,4,5,6,7,8,9,0)$* dans le champ du filtre et choisir la clé de tri jugée pertinente.

Dans cet exemple, on note deux erreurs ou un chiffre ( 0 et 1) prend la place d'une lettre.

Problème 3 : Erreurs d'orthographe

Un dernier problème courant révélé par le lexique est l'orthographe, en particulier pour les textes qui n'ont pas été validés par un correcteur orthographique. C'est un problème mineur si

on entend analyser le corpus avec des outils statistiques puisque ces outils sont surtout sensibles aux mots fréquents. Dans la plupart des cas, un problème d'orthographe est unique et un mot mal orthographié n'apparaît qu'une seule fois dans le corpus.

COMMENT DÉTECTER CE PROBLÈME?

Une manière simple de détecter la présence de problèmes d'orthographe est d'afficher tous les mots qui ne sont présents qu'une seule fois dans le lexique. Les mots dont la fréquence dans le corpus est 1 sont appelés les *hapax*. Pour un corpus assez long, ils constituent à peu près la moitié du lexique. Donc, l'opération peut être assez longue, mais elle est quand même plus rapide que de passer en revue tous les mots du lexique.

LEXIQUE AFFICHER $*freqtot=1 TRI Alphabet

| Fréqtot | |
|---|---|
| 1 | 0raisons |
| 1 | analyses |
| 1 | analyste |
| 1 | certaineméthode |
| 1 | changer |
| 1 | changerai |
| 1 | critiques |
| 1 | d1scours |
| 1 | distinge |
| 1 | divisionsplus |
| 1 | élocence |
| 1 | excellence |
| 1 | genrre |
| 1 | manger |
| 1 | mangerai |
| 1 | politicien |

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Lexique* dans le menu de gauche de l'interface avancée;
2) Choisir *Afficher;*
3) Ensuite, entrer *$*freqtot=1* dans le champ du filtre et choisir la clé de tri jugée pertinente.

Dans cet exemple, on notera la présence d'erreurs sur les mots *0raisons* (oraisons), *d1scours* (discours), *certaineméthode* (certaine méthode), *divisionsplus* (divisions plus), *genrre* (genre), *distinge* (distingue), *élocence* (éloquence).

Une autre solution pour dépister les fautes, consiste à utiliser le scénario de catégorisation grammaticale du lexique que l'on retrouve sous l'onglet *Tâches* dans l'interface intégral de SATO. Sous l'onglet *Tâches*, on choisira l'option *Catégorisation grammaticale* qui offre différents outils pour diverses langues. Pour notre propos, on choisit l'option *Scénario gramr.csa* pour le français.

Le scénario définit une propriété *Gramr,* une propriété symbolique pour le lexique. Les mots qui n'auront pas été reconnus auront la valeur *nil* pour la propriété *Gramr*. On peut faire afficher tous les mots non catégorisés. Parmi ces mots, on retrouvera des noms propres, des sigles, des mots nouveaux et en langue étrangère, mais aussi des mots mal orthographiés.

LEXIQUE AFFICHER $*gramr=nil TRI Alphabet

```
Fréqtot  Gramr
1        nil   0raisons
1        nil   certaineméthode
1        nil   d1scours
1        nil   distinge
1        nil   divisionsplus
1        nil   élocence
1        nil   genrre
```

Pour obtenir ce résultat à partir de l'interface Web, on doit :

1) Cliquer sur le lien *Lexique* dans le menu de gauche de l'interface avancée;
2) Choisir *Afficher;*
3) Ensuite, mettre *$*gramr=nil* dans la champ *filtre;*
4) Finalement, choisir *alphabet* comme propriété de tri.

## 2.5 L'enrichissement lexical

Comme expliqué dans la section *Principes et définitions* de ce guide, SATO construit le lexique des mots du texte selon les règles des alphabets utilisés dans le corpus. Ces règles ne permettent pas toujours, à elles-seules, de repérer les unités terminologiques de la langue.

Il faut savoir, cependant, que les analyseurs statistiques appliqués aux textes sont peu sensibles aux subtilités de la langue. Ainsi, on a pu constaté que l'analyse statistique appliquée aux mots tels qu'ils se présentent, par exemple les formes conjuguées du verbe, donne des résultats assez semblables à l'analyse de textes *lemmatisés*, c'est-à-dire des textes dans lesquels les formes conjuguées du verbe, par exemple, auraient été remplacées par la forme infinitive que l'on retrouve dans les dictionnaires. En fait, comme c'est toujours le cas en analyse de texte, ces décisions de normalisation dépendent de nos objectifs de recherche et de la nature des corpus analysés.

SATO permet à l'analyste du texte de partir d'un texte brut et d'effectuer, si nécessaire, diverses opérations de consolidation terminologique afin de produire des versions enrichies du corpus et de son lexique. La section *Tâches* de l'interface avancée de SATO fournit divers outils pour ce faire. Les formulaires de l'interface SATO documentent chacune de ces tâches. Les paragraphes qui suivent ont donc surtout pour objectif d'indiquer en quoi ces tâches

peuvent être invoquées pour produire des versions enrichies du corpus, si jugé pertinent pour l'analyse et l'interprétation.

### Catégorisation grammaticale

La tâche *Catégorisation grammaticale* a déjà été évoquée pour la vérification orthographique du texte. Deux types d'outils sont disponibles pour cette tâche. Il s'agit de la catégorisation hors contexte, qui s'intéresse au lexique du corpus et aux catégories grammaticales possibles des mots, telles qu'inscrites dans les dictionnaires. On a aussi des catégorisateurs en contexte, généralement de nature probabiliste, qui déterminent la fonction de chacune des occurrences des formes lexicales dans leur contexte syntaxique. Les catégories grammaticales utilisées par ces analyseurs statistiques sont souvent différentes de celles que l'on retrouve dans les dictionnaires, mais elles ont l'avantage d'être spécifiques à chacun des mots dans le contexte de la phrase. Les résultats de la catégorisation grammaticale pourront être mis à profits pour l'enrichissement lexical.

### Locutions et mots complexes

La rubrique *Locutions et mots complexes* de la section *Tâches* de l'interface avancée de SATO fournit des outils pour repérer des mots composés, des abréviations et des locutions.

Comme nous l'indiquions dans la section *Comment SATO peut-il reconnaître les mots ayant une forme complexe : mots composés, sigles, expressions, etc.?,* il est possible d'indiquer à SATO de considérer des expressions comme étant des mots et de les consigner telles quelles dans le lexique du corpus. Mais, pour ce faire, il est préférable d'utiliser les outils d'analyse de SATO plutôt que de marquer manuellement les expressions à lexicaliser.

Diverses stratégies sont proposées par SATO dans la tâche *Dépistage des locutions et des termes complexes* dont voici le sommaire.

1. Introduction
2. Dépistage des mots composés
3. Créer un fichier de locutions
4. Modifier un fichier de locution
5. Appliquer un fichier de locutions
6. Dépistage de locutions par patrons syntaxiques
7. Annuler les liaisons

### Mots composés

Le dépistage de mots composés propose de lier les mots séparés par des traits d'union. Les constructions syntaxique de type interrogatif (*m'aime-t-il?*) ou démonstrative (*cette table-ci*) ne seront pas liées. Cependant, les pronoms du type *celle-ci* seront liés.

Il est à noter que certains mots composés échapperont à cet algorithme. Par exemple, *rendez-vous* ne sera pas lié parce qu'il peut aussi se retrouver dans une phrase interrogative

comme dans *Rendez-vous vos armes?* Il sera toujours possible de forcer la liaison en ajoutant le mot à une liste de locutions, tel qu'expliqué dans *Créer un fichier de locutions*. Le scénario *mot-comp.csa*, que l'on pourra consulter dans la librairie de scénarios SATO, peut aussi en faire trop! Par exemple, le scénario lie des mots composés contenant des nombres comme dans *guerre 14-18*. Mais il liera aussi pp.*12-16*. Dépendant de nos objectifs d'analyse, ces décisions peuvent ne pas convenir. Mais, avec SATO, il est toujours possible de délier des expressions parasites par les opérateurs de catégorisation de la commande *Contexte appliquer*. L'analyste pourra aussi choisir de faire sa propre version du scénario *mot-comp.csa* en s'inspirant du scénario en librairie!

**Fichier de locutions**

Mais, pour des fins d'analyse, on pourra vouloir se concentrer sur des expressions vraiment pertinentes pour nos questions de recherche. C'est le cas d'*assemblée nationale* dans une analyse de corpus portant sur le nationalisme et où l'adjectif *national* est pertinent, sauf lorsqu'il fait partie de dénominations. On peut donc faire une liste des expressions à lier et appliquer ce fichier de locutions pour des fins d'analyse particulières.

Dans SATO, le fichier de locutions est en fait un ensemble de filtres SATO qui seront transformées en commandes *Contexte appliquer* avec opérateurs de catégorisation. Donc, on peut utiliser des caractères de cache pour décrire les mots individuels composant les expressions. Par, exemple *mouvement$ nationa$* dépistera *mouvement national*, *mouvements nationalistes*, etc.

**Attention.** Dans un fichier de locutions, chaque terme de la locution doit être séparé du suivant par un espace. Les mots devraient aussi être en minuscules. Par exemple, si on veut lier une abréviation comme *M.*, on trouvera dans le fichier de locutions la ligne suivante :

```
m*Édition=maj .*Édition=collé
```

m désigne la lettre *m* en minuscule dans le lexique, mais présentée en majuscule dans le texte. Le point sera ici considéré comme un point d'abréviation à condition qu'il soit collé sur la lettre *m*. Dans l'expression on notera qu'il y a un espace après *maj*.

**Exportation du corpus**

Comme illustré dans le chapitre *Processus de catégorisation en contexte pas-à-pas* du Manuel SATO, l'exportation d'un corpus en format SATO permet de produire un fichier textuel qui contient des balises traduisant en clair l'annotation réalisée sur le corpus au moyen de SATO. Le dépistage des locutions, des mots composés et des abréviations fait partie du travail d'annotation. Le corpus exporté dans un nouveau fichier texte portant l'extension *.sat* pourra être soumis à SATO qui mettra alors dans le lexique du corpus les expressions comme **(assemblée nationale*)* repérées au préalable par la fonction *Appliquer un fichier de locutions* de la tâche *Locutions et mots complexes*.

Il est possible d'exporter un corpus annoté dans divers formats, notamment le format XML-TEI servant de langage pivot pour convertir le corpus pour traitement par divers logiciels.