INTEX

INTEX est un logiciel développé au sein du LADL (Laboratoire d'automatique documentaire et linguistique – fondé par M. Gross en 1967). C'est un environnement de développement utilisé pour construire des descriptions formalisées à large couverture des langues naturelles, et les appliquer à des textes de taille importante en temps réel. Les descriptions linguistiques sont formalisées sous la forme de dictionnaires électroniques, de grammaires représentées par des graphes à états finis et de lexiques-grammaires.
INTEX fournit des outils pour décrire la morphologie flexionnelle et dérivationelle, la variation orthographique et terminologique, le vocabulaire (mots simples, mots composés et expressions figées), les phénomènes semi-figés (entre lexique et syntaxe : grammaires locales, description des accords), et la syntaxe (grammaires syntagmatiques).
Il est également utilisé comme **traitement de corpus** : il permet de traiter en temps réel des textes de taille très importante (plusieurs centaines de mégaoctets : une année d'un quotidien, 150 livres de poche). Les opérations typiques sur les textes incluent l'indexation de motifs morpho-syntaxiques, d'expressions figées ou semi-figées, de concordances lemmatisées et l'étude statistique des résultats.

Caractéristique : tous les objets traités (textes, dictionnaires, grammaires) sont à un moment ou à un autre représentés par des **transducteurs à états finis** (= un graphe qui représente un ensemble de séquences en entrée, et leur associe des séquences produites en sortie). Les opérations effectuées sur les textes, dictionnaires ou grammaires peuvent être ramenées à une trentaine d'opérations élémentaires.
Tous les objets sont modifiables. On peut en créer de nouveaux (dictionnaires, grammaires) selon les besoins de l'utilisateur.

Manupulation :

1. Lancer INTEX. Choisir la langue de travail (déterminant l'ensemble des ressources linguistiques à utiliser).
2. Charger un texte ASCII brut (.txt) : (texte de démo : la femme de trente ans (Balzac).

Préparation du texte : se fait en trois passes
1. Le découpage des phrases du texte, reconnaissance des délimiteurs et remplacement des délimiteurs par des étiquettes (ex. {S}). Voir le graphe sentence.fst (FSGraph —> open).
2. Passage du dictionnaire de normalisation des séquences (mots composés) non ambiguës (ex. parce que, aujourd'hui.). Remplacement par des étiquettes. Voir les graphes de normalisation.
3. Deuxième dictionnaire de normalisation qui remplace les élisions, contractions (ne pas), abréviations par des étiquettes. Voir les graphes de preprocessing —> replace.

Des dictionnaires par défaut sont appliqués et le FST du texte est construit.

Résultats du preprocessing : afficher le texte avec tags.
1) liste des formes (tokens)
2) listes des mots trouvés par les dictionnaires par défaut
3) graphe(s) du texte.

Appliquer les ressources linguistiques : (Text → apply lexical resources)
On peut appliquer de nouveaux dictionnaires, de façon à reconnaître, donc à étiqueter, le maximum de formes. La mise à jour des résultats se fait automatiquement. Le résultat des dictionnaires correspond à l'union des dictionnaires utilisés. Toutes ces fenêtres de résultats sont indépendantes des fichiers de ressources du système ; ce sont des fichiers associés au texte traité.

Désambiguation : (Text —> Disambiguation)
Utilisation de grammaires locales pour lever les ambiguïtés lexicales et lexico-syntaxiques. Les ambiguïtés, c'est-à-dire plusieurs hypothèses pour un mot, sont représentées par des graphes ou des expressions rationnelles (régulières). Le but est comme toujours d'obtenir le maximum d'étiquettes, et le minimum de formes restantes (toujours ambiguës ou non reconnues). Les résultats sont mis à jour automatiquement.

Recherche de séquences dans le texte : (Text —> locate pattern)
En utilisant des expressions régulières, par exemple <V :F>, on obtient toutes les formes verbales au future apparaissant dans le texte. Si le nombre de résultats est faible, préférer la construction d'un concordancier qui présente les formes en contexte ; sinon, utiliser l'option « full text » qui présente les formes relevées en couleur.
On peut également choisir de créer son propre graphe (représentant une séquence prototypique) pour mener cette recherche.