ENTRE SENS ET TEXTE, UNE MODÉLISATION

Parler, c'est transformer ce que l'on veut dire en ce que l'on dit, autrement dit un sens en un texte. Pour décrire cette correspondance entre sens et textes, les graphes sont des outils privilégiés.

La linguistique

On appelle sujet parlant ou locuteur quelqu'un qui parle, c'est-à-dire quelqu'un qui cherche à communiquer avec d'autres gens en produisant des paroles ou un texte écrit. Décrire ce que produit un locuteur et la façon dont il le produit est l'objet d'une science, la linguistique, elle-même reliée à d'autres sciences comme la psychologie, la logique ou les sciences cognitives.

La sémantique s'occupe du sens des mots et des textes alors que la syntaxe s'occupe de leurs constructions. Ainsi, la phrase "d'incolores idées vertes dorment furieusement" est syntaxiquement correcte, alors que "les enfants joue dans la cour" ne l'est pas. Du point de vue sémantique, c'est le contraire.

Normalement, quand quelq'un parle, c'est qu'il a quelque chose à dire. Parler, c'est transformer un sens (ce qu'on veut dire) en un texte (ce qu'on dit). Modéliser une langue, c'est décrire la correspondance entre les sens et les textes de cette langue.

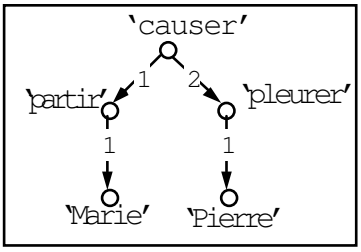
Qu'est-ce que le sens ?

Avant de se demander comment décrire une telle correspondance, il faut savoir comment décrire textes et sens. Pour les textes, c'est relativement simple. Il s'agit de suites de mots avec en plus, dans le cas de l'oral, le ton, l'intonation, l'accent, etc. (on dit la prosodie). Mais qu'est-ce que le sens associé à un texte et comment le représenter ? Nous présentons ici un système de représentation basé sur des graphes. L'idée est que le sens d'une phrase naît de l'interaction des sens des mots. Cette interaction peut être visualisée au moyen d'un graphe.

Certains mots agissent sur le sens d'autres mots et les mettent en relation. Ces mots sur lesquels un mot agit ainsi sont appelés ses arguments sémantiques. Par exemple, dans la phrase :

(1) "Pierre pleure à cause du départ de Marie",

le verbe "pleure" a comme argument sémantique "Pierre" et c'est le seul. On dit également que "pleure" est une prédication sur "Pierre". Syntaxiquement parlant, "Pierre" est le sujet de "pleure". Pour continuer sur cet exemple, le groupe "à cause de" a deux arguments : "départ" et "pleure". Nous pouvons continuer l'étude de ces relations ainsi et regrouper tout cela sous forme d'un graphe orienté :



Dans ce graphe, chaque arête représente la relation entre le sens d'un mot prédicat et l'un de ses arguments. Nous pouvons remarquer que ce graphe est connexe (voir lexique), c'est-à-dire que tout mot est relié à tous les autres par une suite de relations (d'arêtes). Ceci assure l'unité de sens de la phrase.

Les étiquettes sur les arêtes du graphe indiquent le numéro de l'argument et permettent de distinguer les différents arguments d'un même prédicat. D'autres phrases ont le même graphe et sont donc des paraphrases de cette phrase initiale :

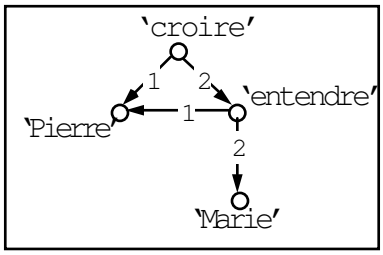
(2) "Pierre pleure, parce que Marie part",

(3) "Les pleurs de Pierre sont dus au départ de Marie",

(4) "Le départ de Marie fait pleurer Pierre".

Comme on le voit le sens 'causer' peut être exprimé en français de multiples façons : la préposition "à cause de", la

conjonction de subordination “parce que”, la tournure verbale “être dû à” ou la construction dite causative avec “faire“ suivi d'un verbe à l’infinitif. Le graphe sémantique d'une phrase peut avoir des cycles. C'est le cas de la phrase “Pierre croit entendre Marie” ou de la paraphrase “Pierre croit qu'il entend Marie” (quand “il“ renvoit à “Pierre”).



Insistons sur le fait qu'un graphe sémantique ne représente qu'une partie du sens, le sens situationnel qui renvoie à la situation que l'on veut décrire. A cela s'ajoute entre autres une structuration communicative qui indique la façon dont on veut communiquer l'information en fonction de ce dont on est en train de parler (le thème), de ce qu'on veut communiquer (le rhème) et d'autres choses. C'est la différence de structuration communicative sous-jacente qui fait que les phrases ci-dessus, bien que décrivant la même situation, ne peuvent être employées dans les mêmes contextes.

Passer du sens au texte

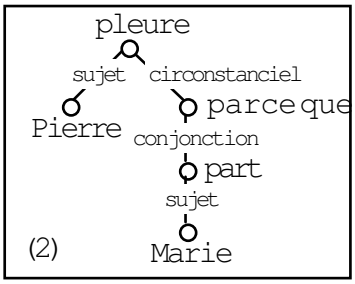
Le passage du sens au texte peut être modélisé par un ensemble fini de règles qui permettent d'associer un graphe sémantique comme celui de la page ci-contre à une suite de mots. Une partie de cet ensemble de règles forme le lexique et concerne la description des mots. Par exemple, “pleurer” est un verbe du premier groupe et a un unique argument, alors que “donner” a trois arguments qui se réalisent comme sujet, objet direct et objet indirect (“Pierre donne un cadeau à Marie”). L'autre partie forme la grammaire proprement dite à savoir les règles de conjugaison ou le fait que le sujet se place avant le verbe, que le complément d'objet indirect est introduit par la préposition “à” ou qu'il se pronominalise par le pronom “lui” placé devant le verbe (“Pierre lui donne un cadeau”).

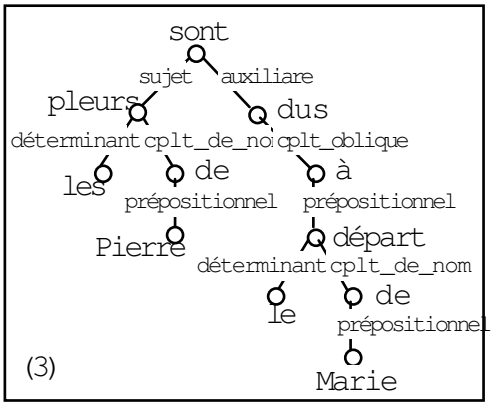
Pour décrire la correspondance entre graphes sémantiques et textes, il est préférable de considérer des représentations intermédiaires, notamment une représentation dite syntaxique. La grammaire devient alors modulaire. Un premier module assure la correspondance entre le graphe sémantique et la représentation syntaxique et un autre module assure la correspondance entre la représentation syntaxique et le texte.

La représentation syntaxique que nous adoptons s'appelle un arbre de dépendance syntaxique (voir lexique). Ses nœuds sont les mots de la phrase et ses arcs sont étiquetés par des relations syntaxiques indiquant la fonction du mot dépendant par rapport au mot dont il dépend. Voici les arbres syntaxiques des phrases :

(2) “Pierre pleure parce que Marie part”,

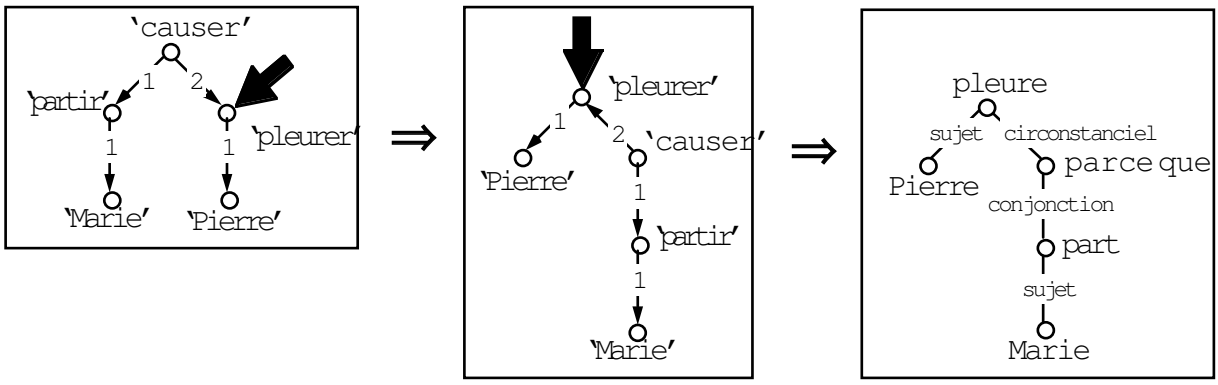
(3) “Les pleurs de Pierre sont dus au départ de Marie”,





Remarquez qu'un mot ne correspond pas forcément à une chaîne de caractères entre deux espaces. Ainsi, “au” est considéré comme l'amalgame de deux mots, “à” et “le”, tandis que “parce que” est considéré comme un seul mot.

Pour passer du graphe sémantique à un arbre de dépendance, il faut choisir le nœud qui correspondra à la racine de l'arbre (ce choix est guidé par la structuration communicative), puis suspendre le graphe par cette racine (voir dessin page suivante). Ensuite, il faut lexicaliser les sens, c'est-à-dire choisir pour chaque sens dans le graphe un ou plusieurs mots qui l'expriment (comme par exemple exprimer “causer” par “à cause de” ou “être dû à”).
En français, comme dans la plupart des



Suspension de la phrase à "pleurer"

langues, la racine de l'arbre de dépendance doit être un verbe (qu'on appelle le verbe principal). Par exemple, pour l'arbre de (2), le nœud sémantique 'pleurer' est choisi pour exprimer la racine de l'arbre et donne le verbe "pleure" ; le sens 'causer' devient un dépendant de ce verbe et est exprimé par la conjonction de subordination "parce que". Pour l'arbre de (4), le nœud 'causer' est choisi pour exprimer la racine de l'arbre et donne la configuration "sont dus à" ; le sens 'pleurer' devient un dépendant de ce verbe et est exprimé par le nom "pleurs".

S'il y a des cycles dans le graphe, il faut les couper soit sur un arc (ce qui donne "Pierre croit entendre Marie" pour notre deuxième exemple de graphe), soit sur un nœud. De cette opération résulte l'introduction de pronoms, comme le pronom "il" renvoyant à Pierre dans "Pierre croit qu'il entend Marie".

Le passage de l'arbre de dépendance syntaxique au texte consiste à ordonner les nœuds de l'arbre syntaxique. Les règles d'ordre sont grosso modo des règles qui indiquent comment un nœud se place par rapport au nœud dont il dépend selon le type de l'arc qui les relient. Nous n'en parlerons pas plus ici.

Comme on le voit, l'activité du langage peut être modélisée par des manipulations de structures mathématiques, comme des graphes, des arbres ou des suites. On retiendra que la particularité du signal linguistique (que ce soit de la parole ou un texte écrit) est d'être linéaire. Quand on parle, on ne peut que produire des mots les uns après les autres. Par contre, il y a tout lieu de penser que le sens, comme notre cerveau, est un objet multidimentionnel. Le passage du sens au texte comprend deux étapes essentielles : la hiérarchisation, c'est-à-dire le passage d'un graphe sémantique à un arbre syntaxique (un arbre est un graphe hiérarchisé), et la linéarisation, c'est-à-dire le passage d'un arbre syntaxique à une suite de mots.

Sylvain Kahane

Igor Melčuk

La présentation faite dans cet article de la modélisation linguistique est basée sur les travaux d'Igor Melčuk, l'un des pionniers de la modélisation formelle des langues.

Igor Melčuk est né le 19 octobre 1932 à Odessa en Ukraine. Il a étudié puis travaillé dans l'ex-URSS jusqu'en 1976. Il a alors été expulsé pour des raisons politiques, il est devenu depuis citoyen canadien et travaille toujours à l'Université de Montréal.

La théorie de Melčuk, la Théorie Sens-Texte, est né dans les années 60 à Moscou. Elle est exposée dans son livre *Dependency Syntax* (1988, SUNY) ou dans le cours qu'il a donné au Collège de France en 1997. Les arbres de dépendance, utilisés comme représentation syntaxique dans cette présentation, ont été développé par Lucien Tesnière, qui restera comme l'un des linguistes majeurs du 20ème siècle.

Lucien Tesnière est né en 1893 à Mont-Saint-Aignan et mort en 1954 à Montpellier où il enseignait la linguistique aux élèves de l'Ecole Normale d'instituteurs. Il est le premier à avoir développé une théorie syntaxique entièrement basée sur la dépendance, bien qu'on trouve déjà des dessins d'arbres de dépendance dans des grammaires du 19ème siècle. Resté marginal de son vivant, il n'a laissé qu'un ouvrage, *Eléments de syntaxe structurale*, publié à titre posthume en 1959 (Klincksieck, réédité) et devenu depuis un classique.