Étiquetage semi-automatique de la prosodie dans les corpus oraux : algorithmes et méthodologie

Estelle Campione

Thèse dirigée par Jean Véronis, soutenue le 08 décembre 2001

L'objectif de cette thèse est de mettre en évidence la faisabilité de la transcription prosodique de grands corpus, à l'aide d'outils semi-automatiques destinés à (1) réduire le coût humain de la transcription, (2) lui apporter une base objective. En effet, alors que l'on dispose désormais pour l'écrit de corpus de dizaines, voire de centaines de millions de mots, les corpus oraux transcrits avec des indications prosodiques sont rares et dispersés. La prosodie semble pourtant indispensable à l'interprétation des transcriptions d'oral, et surtout à l'étude des relations qu'elle entretient avec les phénomènes syntaxiques et pragmatiques.

Nous proposons un codage prosodique « large » qui délimite seulement les unités majeures, dans une perspective de lisibilité et d'exploitation de grands corpus pour l'analyse des phénomènes syntaxiques et pragmatiques. Notre travail est sous-tendu par la volonté constante de limiter les présupposés théoriques, de façon à produire des corpus annotés les plus neutres possibles vis-à-vis des théories particulières, et nous avons fait le choix de transcrire seulement les événements qui correspondent à une réalité objective et qui sont par conséquent repérables par l'instrumentation tels que les proéminences mélodiques, les pauses silencieuses, les « marques du travail de formulation » (*euh* et allongements syllabiques), *etc.* L'exemple ci-dessous montre un fragment de corpus transcrit avec notre système. La notation est volontairement assez simple : les énoncés sont découpés en segments qui apparaissent chacun sur une ligne séparée, précédée de la position temporelle en secondes à partir du début de l'enregistrement ; les montées et descentes intonatives importantes sont notées par les flèches (↗ et ↘), les allongements syllabiques d'hésitation sont marqués par deux-points (:) ; les tirets (- et --) marquent deux niveaux de pause.

0.0 on on dit il y a une anecdote qui dit pourquoi: - Corton-Charlemagne ↘
4.5 -
5.1 ben parce que le: Charlemagne euh - paraît-il ↗ ne buvait que des rouges ↗
9.0 -
9.5 parce qu'il voulait pas ↗
10.4 -
10.9 il il se tachait ↗
12.0 -
12.6 sa: sa: - il ne ne buvait que* des Blancs pardon ↗
15.9 -
16.3 euh parce qu'il ne voulait pas ↗ se tacher ↗
17.9 -
18.2 sa: sa barbe ↗
19.7 -
20.2 et donc euh - c'est pour ça qu'un jour ^ les habitants de: - d'Aloxe-Corton: - ont dit ↗ mais alors →

A l'aide d'un corpus de travail constitué de près de six heures de parole (lue et spontanée) dans cinq langues (anglais, allemand, espagnol, français et italien), que nous analysons avec des outils informatiques et statistiques, nous proposons dans un premier temps une analyse rigoureuse et systématique des différents indices et de leur interaction. Nous étudions ainsi près de 60000 mouvements mélodiques et 6000 pauses, et l'influence de facteurs tels que la langue, le sexe, le débit ou le type de parole. Si nos résultats confirment le plus souvent, en les complétant, les études précédemment publiées, ils apportent parfois un point de vue nouveau, voire contredisent certaines données tenues pour acquises.

Dans un deuxième temps, nous proposons un algorithme qui prend en compte l'ensemble des contraintes et interactions pour produire en sortie un codage prosodique des corpus. Nous avons découplé de façon soigneuse d'une part la phase d'extraction d'indices, automatique ou manuelle, et d'autre part l'interprétation de ces indices à l'aide de règles et contraintes. Cette approche modulaire permet une meilleure compréhension du rôle des différents indices et un travail de mise au point plus aisé. Elle permet aussi une certaine indépendance de la langue : les modules de bas niveau sont applicables à diverses langues et ont été testés sur les cinq langues de notre corpus, et seul le système de règles d'interprétation est spécifique à chaque langue (dans le cadre de ce travail, nous proposons un tel système seulement pour le français).

Étude critique d'un remarqueur de la fin du XVIIe siècle : Nicolas Andry de Boisregard

Magali Seijido

Thèse dirigée par Claire Blanche-Benveniste, soutenue le 08 décembre 2001

Cette thèse est une étude critique d'un auteur peu connu - Nicolas Andry de Boisregard - qui a publié deux recueils de remarques complémentaires : les Réflexions sur l'usage présent de la Langue Françoise ou Remarques nouvelles et critiques touchant la politesse du langage édité en 1689 et la Suite des Réflexions Critiques sur l'usage présent de la Langue Françoise édité en 1693.

Le genre des Remarques et des Observations sur la langue du XVIIe siècle a été pendant longtemps peu étudié, pourtant il constitue un contrepoint intéressant aux grammaires traditionnelles : il a tendance à être axé sur des données plutôt que sur des théories. Les « remarqueurs » traitent des subtilités et des idiotismes de l'usage, de faits de langue dans le détail plutôt que dans l'ensemble ; ils portent un intérêt particulier aux usages douteux, à la variation, à l'instabilité et au changement de la langue. Leurs témoignages sont donc très précieux pour l'histoire du français et de sa standardisation. Cependant, pour un lecteur moderne, l'accès à ce contenu est parfois rendu difficile par la forme même de ces ouvrages : ils se présentent généralement comme des recueils de courtes remarques détachées. Il n'y a pas de lien logique entre les articles, et sans index, il est peu aisé d'y rechercher un phénomène précis ou de les exploiter pour l'étude d'un problème grammatical.

Le premier objectif de cette thèse est de proposer des outils qui permettraient de faciliter la consultation des Réflexions (1689) et de la Suite des réflexions (1693) et qui pourraient constituer un matériau pour l'histoire de la grammaire, pour l'étude de la variation et de la grammaticalisation. J'ai réalisé un classement moderne des remarques en cinq rubriques : prononciation et orthographe, grammaire, lexique, rhétorique et règles de bienséance, qui aboutit à une nouvelle présentation : j'ai « scanné » les remarques des deux recueils et je les ai reproduites en suivant cette réorganisation.

Je propose une présentation très générale des remarques grammaticales et une étude plus détaillée du verbe et des problèmes de cohésion syntaxique. En grammaire, la majorité des remarques aborde des questions liées à la morphosyntaxe des catégories. Toutes les catégories sont représentées : déterminant, pronom, substantif, adjectif, adverbe, préposition, connecteur. Andry fait de nombreuses observations sur le verbe et le groupe verbal ; il s'intéresse principalement aux valences mais il fait également des remarques sur la morphologie, l'accord, les modes et les temps.

Bien que ce classement ait demandé un travail de catégorisation, il n'a pas pour objectif d'aboutir à une grammaire organisée. Les « remarqueurs » et les grammairiens n'ont pas la même démarche. Andry, en refusant de faire une grammaire, en refusant l'ordre fondé sur les parties du discours, ne propose pas de théorie générale sur l'ensemble de la langue. Deux sortes d'erreurs d'interprétation consisteraient à prêter aux « remarqueurs » des théories qui n'ont pas lieu d'être ou à les qualifier de descriptivistes comme l'a fait Chomsky à propos de Vaugelas. Le rôle d'Andry est de participer concrètement à la standardisation du français en observant et en commentant principalement les usages qui posent problèmes et la variation. A titre d'exemple, si l'on prend le traitement qu'il fait des valences verbales, il ne s'intéresse pas aux principes généraux : la majorité de ses remarques portent sur des verbes qui ont plusieurs constructions concurrentes en usage. Il adopte alors plusieurs positions. La position la moins fréquente consiste à sélectionner une construction et à condamner l'autre. Dans d'autres cas, il accepte la « variation » et considère les deux constructions comme strictement équivalentes. Les constructions concurrentes peuvent également différer par une nuance de style mais le plus souvent il les distingue par le sens : elles peuvent correspondre à une simple opposition de traits (+ /- animé, + /- fort, + /- humain) ou à deux sens différents.

Dans cette thèse, je propose également des hypothèses sur l'évolution de la définition du bon usage au XVIIème siècle. J'interviens contre la simplification qui associe, sous le règne de Louis XIV, le bon usage à l'usage de la cour. Un lecteur aussi attentif que R. A. Lodge note que « au cours des cinquante-cinq ans du règne absolu de Louis XIV, la concentration du pouvoir dans la personne du roi assura le maintien du rôle de la cour comme garant du bon usage. » Or, Andry observe peu les usages en fonction de ce que nous appellerions « la stratification sociale ». La cour ne sert de modèle que pour les règles de prononciation. Pour discuter de la norme grammaticale, il se réfère principalement à des productions écrites : plus de cent ouvrages. Plusieurs types de productions sont représentés : le

domaine technique, le domaine poétique et le domaine de la littérature en prose. Un phénomène est tout à fait remarquable : il accorde une place centrale à la traduction ; d'une manière générale, elle tend à remplacer la cour en matière de bon usage.