Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Dimitrios MELETIS. The Nature of Writing : A Theory of Grapholinguistics. *Fluxus Editions* (coll. « Grapholinguistics and its applications », vol. 3). 2020. IX, 461 pages. ISBN : 978-2-9570549-2-3.

Lu par **Pascal Vaillant**

Université Sorbonne Paris Nord (Paris 13), LIMICS

Il est très difficile de définir les concepts d'une linguistique de l'écrit en s'abstrayant des spécificités d'une langue ou même d'une écriture, tant celles-ci diffèrent dans leurs unités et leurs principes combinatoires. L'ouvrage The Nature of Writing : A Theory of Grapholinguistics, *de Dimitrios Meletis, est la tentative la plus aboutie dans cette direction.*

Dimitrios Meletis, universitaire autrichien, livre avec *The Nature of Writing* une somme théorique impressionnante sur la langue écrite. L'ambition de Meletis est tout à la fois de définir un cadre sémiotique général de l'écriture (partie *Description*) et de fournir des guides d'application permettant d'analyser, dans ce cadre, la manière dont les utilisateurs des langues font usage des systèmes d'écriture et la manière dont ces derniers « s'adaptent » aux langues qu'elles représentent (partie *Explanation*). Il s'y est appliqué avec une rigueur impressionnante et a creusé en profondeur chaque concept et chaque outil méthodologique, jusqu'à obtenir un système théorique cohérent.

Définition du cadre

La première partie de l'ouvrage s'applique à fournir un cadre descriptif unifié à la linguistique de l'écrit (l'auteur utilise le terme *grapholinguistics*, calqué de l'allemand *Schriftlinguistik*[1]), qui puisse ensuite être appliqué à l'analyse de plusieurs langues écrites en utilisant des concepts communs. L'écrit est ici défini comme lié à une langue : les tentatives de sémasiographie pure (comme le *Bliss*) n'entrent pas dans le champ d'étude. Il n'est pas pour autant une simple transposition de la langue orale et « possède une vie propre » : en témoigne la

1 Dimitrios Meletis est issu du monde académique germanophone, où s'est développée depuis quatre décennies une tradition forte de linguistique de l'écrit.

présence, dans presque toutes les écritures, d'unités (les signes de ponctuation) ou d'oppositions (entre majuscules et minuscules) n'ayant pas d'équivalent dans la langue parlée.

L'auteur a voulu, dans cette première partie, élaborer des concepts génériques à partir d'une masse de travaux antérieurs qui ne formaient pas un tableau cohérent. Il ne l'a pas fait en réinventant la roue : l'abondance de ses sources montre une très grande culture du domaine, qu'il fait partager au lecteur en en dressant l'état de l'art. Il a en revanche constaté que les travaux antérieurs ne convergeaient pas vers des définitions superposables de concepts aussi fondamentaux que ceux de graphème ou d'unité élémentaire de l'écrit (le premier n'étant d'ailleurs pas, dans tous les travaux, défini comme la seconde). Si les réponses à ces questions sont différentes, c'est souvent que les questions n'ont pas été formulées de la même manière, ni avec les mêmes présupposés : alors que certains ont cherché à définir la structure d'une écriture comme un système de signes graphiques et sont partis des formes pour étudier leur combinatoire, puis leurs fonctions, d'autres l'ont abordée avant tout comme un système de transcription d'une langue orale et sont partis d'une typologie des unités linguistiques pour étudier les différentes manières de les représenter. Ces deux approches conduisent à des unités parfois complètement différentes. En outre, les travaux sur l'écriture, à part les ouvrages à caractère historique comme ceux de Marcel Cohen ou d'Ignace J. Gelb, sont rarement informés de la diversité des écritures du monde, comme l'auteur le note en préambule : « *descriptions of writing systems coexist but rarely reference one another* ». Or la conception que l'on peut se faire du fonctionnement de l'écrit diffère largement selon que l'on décrit l'écriture du chinois ou celle de l'allemand. L'impressionnante culture linguistique de Meletis en a fait l'un des rares à pouvoir réfléchir en profondeur aux questions interlinguistiques sans se contenter de généralités ou de ouï-dire, comme on en voit souvent (notamment concernant le chinois). Le premier apport de l'ouvrage est donc une cartographie du domaine étudié, la langue écrite, qui catégorise trois grands modules au sein desquels les objets d'étude sont différents : la graphétique, la graphématique et l'orthographe.

La *graphétique* décrit la combinatoire interne et externe des unités d'un système d'écriture indépendamment de la langue (dans ce sens restreint, *script*). Dans toutes les écritures, les segments de base (qu'il s'agisse de lettres ou de caractères) se distinguent les uns des autres par la disposition de formants graphiques plus fondamentaux. La combinatoire de ces formants, dans l'espace « segmental », est l'un des paramètres d'une typologie des écritures. C'est également dans le cadre de la graphétique que se définit l'allographie, qui est la possibilité pour un même élément abstrait (la lettre |a| par exemple) d'apparaître sous plusieurs formes (A, a, ɑ)[2]. Au-delà de l'espace segmental, la graphétique caractérise également la manière dont les unités se combinent dans l'espace de la ligne, puis dans l'espace de la page.

2 Certaines de ces variations sont régulées par le plan graphématique (comme les variantes positionnelles des lettres arabes ou l'opposition majuscules minuscules) ; d'autres sont « libres » ou plus exactement associées à des valeurs connotatives (comme le choix de la police de caractères). Dans tous les cas leur description relève du module graphétique.

La graphétique est le lieu de la description du matériau graphique de l'écriture ; cependant l'analogie entre graphétique et phonétique (suggérée par leur suffixe commun) a ses limites. En effet, si la phonétique est la description de la pure substance de la langue parlée (la forme lui étant imposée de l'extérieur, par des oppositions qui se manifestent dans les signifiants d'une langue donnée), la graphétique possède en elle-même une opposition forme/substance : que l'alphabet latin serve à noter l'allemand ou le wolof, les différences dans l'espace graphique qui permettent de distinguer la lettre E de la lettre F sont les mêmes.

La *graphématique* établit le lien entre les formes utilisées pour l'écriture et le système linguistique qui lui sert de support. Le concept central de la graphématique est le graphème et l'un des apports majeurs de Meletis est d'en proposer une définition à la fois rigoureuse et généralisable.

Beaucoup de travaux antérieurs ont eu tendance soit à éviter le terme de graphème, soit à l'utiliser pour désigner un élément d'un système d'écriture (en quelque sorte comme un mot à consonance savante pour « lettre »), soit à en proposer une définition *ad hoc* liée à un usage particulier (par exemple pour désigner les segments de chaîne écrite correspondant à des phonèmes, usage fréquent dans le domaine de la pédagogie de l'orthographe). Nina Catach, que l'auteur cite peu (peu de ses sources sont en français) proposait en 1979 une définition générique du graphème (« plus petite unité de la chaîne écrite ayant un correspondant phonique et/ou sémique susceptible d'une analyse linguistique »), mais tout le développement ultérieur de son propos utilisait le terme pour établir une typologie des fonctions des lettres et séquences de lettres en français. Semblablement, comme le note l'auteur, beaucoup de définitions sont spécifiques à une langue et à un système d'écriture. Meletis propose donc de caractériser le graphème par trois critères : (1) c'est un segment distinctif au niveau du sens (distinctif, mais pas nécessairement *uniquement* distinctif, puisqu'il peut correspondre à un morphème, comme c'est le cas général dans le chinois écrit) ; (2) il correspond à une unité fonctionnelle du système linguistique (mais, point important, pas nécessairement une unité du niveau phonologique) ; (3) il doit être minimal (ce qui exclut par exemple |q| de l'inventaire des graphèmes de l'allemand, puisqu'il n'y a pas une paire minimale où |q| apparaît sans |u|) ; le critère de minimalité n'implique pas qu'il s'agisse d'un *segment* minimal dans le découpage linéaire de la chaîne écrite, ce qui permet de rendre compte, par exemple, des jamo coréens. La robustesse de cette définition est éprouvée par son instanciation fructueuse dans toutes les grandes familles de systèmes d'écriture : allemand, thaï, chinois, japonais, coréen.

Au-delà du graphème, ce plan est également celui où s'élaborent les concepts de mot graphématique et de phrase graphématique.

L'*orthographe*, enfin, est le module qui décrit les contraintes d'usage, culturellement et socialement normées, qui s'imposent aux combinaisons d'unités graphématiques. Elles décrivent un sous-ensemble de l'espace des combinaisons possibles.

Évaluer l'adaptation des écritures et des orthographes

La deuxième partie du livre de Meletis est celle où se met en évidence l'utilité du cadre théorique proposé, y compris pour les lecteurs qui ne sont pas fascinés par les classifications *per se*. L'auteur met en effet ses concepts à l'épreuve en les utilisant pour expliquer comment les utilisateurs des langues écrites adaptent les systèmes d'écriture à leurs besoins. Le cadre conceptuel de la théorie de la naturalité (expliqué dans le premier chapitre de cette partie) lui sert à définir des outils méthodologiques permettant d'évaluer différents niveaux d'« adaptation » (*fit*) des écritures aux langues qu'elles représentent. Cette adaptation est mesurée structurellement, au niveau du système linguistique, mais également sur le plan de la performance, ainsi que dans la sphère sociolinguistique (valeurs socialement attribuées à l'écriture et aux différentes façons d'écrire).

Cette partie permet d'éclairer certains débats. Il est, par exemple, notoirement habituel, dans l'espace francophone, de se diviser sur les mérites de l'orthographe. Elle est décrite par certains comme « illogique » (par opposition à des orthographes jugées plus « transparentes », comme celle de l'espagnol). Meletis aide à considérer cette question plus globalement en ajoutant au critère de transparence phonologique celui de transparence morphématique.

Une meilleure compréhension des choix de représentation formelle

L'apport de l'ouvrage pour la linguistique informatique est indirect, du fait que la plupart des questions liées aux particularités de la langue écrite ont déjà, en 2021, trouvé des solutions par défaut. Ainsi, la diversité des unités linguistiques, de leurs niveaux de définition et de leurs règles de combinaisons, a été soigneusement prise en charge par la norme Unicode (qui est bien plus qu'un répertoire de formes graphiques) et les trois décennies de travail de spécialistes qui l'ont produite. Le livre de Meletis permet plutôt de mieux comprendre et d'éclairer rétrospectivement ce travail et d'en apprécier l'ingéniosité (pour ne citer qu'un exemple, dans la représentation du coréen).

La définition des mots, qui passe par une étape graphétique (pour les langues alphabétiques, par exemple, l'extraction des segments séparés par des espaces), puis par une étape graphématique (détermination des cas où certains signes comme l'apostrophe ou le trait d'union font partie du mot, repérage des mots composés, des verbes séparables en allemand…) est en général un choix par défaut intégrés dans l'étape de *tokenization* des chaînes de TAL et que la communauté fait sans y penser, à l'exception notable des travaux sur les langues asiatiques. L'ouvrage de Meletis a le mérite d'en éclairer les mécanismes sous-jacents.

Conclusion

Le livre de Meletis est une somme d'érudition et une impressionnante nouvelle classification cohérente des connaissances sur l'écrit, dans un cadre commun, adaptable à toutes les langues écrites. Et c'est une classification qui n'ignore pas les travaux antérieurs, mais les incorpore et s'en informe. Mais l'intérêt de cet ouvrage ne se limite pas à une belle ontologie de la linguistique de l'écrit. Il fournit également des outils conceptuels pratiques, précieux pour faire avancer, boussole en

main, des discussions qui avaient souvent tendance à rester indécidables, faute d'être ancrées à un socle théorique solide et à des indicateurs bien définis.

L'ouvrage *The Nature of Writing* s'impose d'emblée comme la référence bibliographique majeure de tous les travaux futurs de linguistique de l'écrit.

Vivi NASTASE, Stan SZPAKOWICZ, Preslav NAKOV, Diarmuid Ó SÉAGDHA. Semantic Relations Between Nominals, Second Edition. *Morgan & Claypool publishers*. 2021. 218 pages. ISBN : 9-781-63639-088-8.

Lu par **Yannis HARALAMBOUS**

IMT Atlantique, UMR CNRS 6285 Lab-STICC

Vivi Nastase a soutenu sa thèse à Ottawa, auprès de l'illustre Rada Mihalcea – les deux chercheuses ont un point commun : leur alma mater*, qui est l'université technique de Cluj-Napoca en Roumanie. En 2003 se forme le noyau des auteurs du présent ouvrage, quand Nastase publie avec Stan Szpakowicz (également de l'université d'Ottawa), un article qui porte déjà sur les relations sémantiques. En 2007, Preslav Nakov (de Berkeley) ainsi que trois autres personnes se joignent à eux pour proposer un défi lors de la conférence SemEval, défi qui porte sur le même sujet que cet ouvrage : la classification de relations sémantiques entre nominaux. En 2013, Diarmuid Ó Séaghdga (de Cambridge, UK) se joint à eux et ils transforment les résultats du défi, ainsi que l'expérience accumulée au fil des années, en volume de la collection Synthesis Lectures de Morgan & Claypool. La deuxième édition de celui-ci est l'ouvrage que nous nous proposons de décrire. Sorti en 2021, il a presque doublé de taille (on est passé de 120 à 218 pages). Mis à part quelques ajouts mineurs aux quatre chapitres de la première édition, la véritable contribution de la deuxième édition consiste en l'arrivée d'un cinquième chapitre sur – tendance actuelle oblige – le* deep learning*.*

L'ouvrage comporte cinq chapitres, le premier n'étant qu'une brève mise au point terminologique ainsi qu'une mise en garde sur ce que le livre « n'est pas ». À noter qu'on y trouve une définition du nominal : il ne s'agit pas d'un groupe nominal comme on pourrait le croire mais, plus spécifiquement, d'un nom, d'un nom propre, d'un nom déverbal, d'un nom désadjectival, d'un nom précédé d'un modificateur (l'exemple donné est celui d'un participe passé suivi d'un nom : *processed food*) ou, de manière récursive, d'une suite de nominaux. En effet, là où le français utilise plutôt des groupes prépositionnels, l'anglais va accumuler des noms qui se suivent sans aucune préposition, ainsi, la « couleur de feuille de chêne » sera en anglais une suite de quatre noms : *oak tree leaf color*.

Relations entre nominaux, relations entre concepts

Comme indiqué dans le titre du chapitre 2, l'ouvrage emprunte deux points de vue en parallèle : celui du traitement automatique de la langue (auquel cas on s'intéresse aux nominaux dans les textes) et celui de la représentation des connaissances (auquel cas on s'intéresse aux relations entre concepts, ce qui constitue l'armature même des ontologies, les arêtes de leur structure de graphe).

Après un aperçu historique des interactions entre langue et connaissance, les auteurs enchaînent sur un très intéressant historique des relations sémantiques, d'abord dans les textes (et donc entre nominaux), et ensuite dans les ontologies (et donc entre concepts). Dans la première partie (les textes), il est passionnant de parcourir les tentatives de classification de relations sémantiques, entre 1826 (par l'un des frères Grimm !) et aujourd'hui. Dans la deuxième partie (les ontologies), on s'aperçoit que les problématiques sont les mêmes, même si ceux qui ont proposé des classifications s'intéressent plutôt aux propriétés des relations (du type : faut-il distinguer les méronymies transitives des intransitives ?) qu'à celles des arguments.

Pour mettre un peu d'ordre dans tout cela, les auteurs de l'ouvrage proposent une liste des dimensions de variation des relations sémantiques : l'aspect ontologique ou idiosyncratique (autrement dit : la relation dépend-elle du contexte ?), l'arité de la relation (les auteurs ratent une formidable occasion de parler des graphes conceptuels de Chein & Mugnier), la question de l'ouverture ou fermeture de l'ensemble des relations, l'ordre des relations (second ordre = relations de relations), le domaine et le seuil de précision des relations.

Ce chapitre est vraiment très intéressant pour qui veut avoir une vision globale de la modélisation des relations sémantiques, surtout lorsque l'on se pose des questions générales sur les méthodes et les outils à utiliser, selon le matériau textuel disponible et les objectifs à atteindre.

Extraction supervisée de relations sémantiques

Au chapitre 3 se pose la question de l'extraction *supervisée* de relations sémantiques, dans le sens où les classes de relations sont connues à l'avance et où, idéalement, on a déjà annoté les arguments de relations potentielles. Il commence par un historique des nombreux défis qui ont été posés dans des conférences : MUC (*Message Understanding Conference*, 1987-1997), ACE (*Automatic Content Extraction*, 1999-2008) et Sem-Eval 2007. Ensuite, après une section dédiée aux travaux spécifiques aux arguments de type nom-nom, il est question de certains corpus collaboratifs comme la Wikipédia (et en particulier les info-boîtes), DBpedia, YAGO, WikiNet et Freebase (qui, entre-temps, a été racheté par Google et ensuite injecté dans WikiData), ainsi que de corpus dans des domaines spécialisés, comme la médecine. Après cette partie introductive, on entre dans le vif du sujet : les propriétés (que les datamineurs appellent *features*) et les algorithmes d'apprentissage, où l'on retrouve toutes les méthodes classiques.

Ce chapitre est très bien écrit et très synthétique, il fournit un bon panorama des méthodes classiques (d'avant le *deep learning*) avec énormément de références.

Extraction non supervisée ou semi-supervisée de relations sémantiques

Vu l'extrême variété de relations possibles et l'énorme quantité de données textuelles dont nous disposons aujourd'hui, l'extraction non supervisée – et donc ouverte à tout – est bien plus prometteuse en matière de résultats que l'extraction supervisée. Le chapitre retrace un grand nombre d'approches, à commencer par les approches historiques : extraction de relations à partir d'un dictionnaire en 1981, utilisation de motifs pour extraire la relation d'hyponymie par Hearst en 1992. Ce

travail est très intéressant parce qu'il ne s'est pas limité au cas direct (« X est Y ») mais a exploité des cas indirects comme « X, de même que Y », « X, y compris Y », etc. : dans les deux cas on peut déduire que « Y est un X »). L'approche de Hearst a servi à initier un *bootstrapping* : on trouve des relations, on s'en sert pour former des nouveaux motifs (à trous), et on recommence. Ce faisant, le nombre de relations augmente de manière significative, mais aussi le bruit provenant de mauvaises interprétations, cette source d'erreur est appelée *dérive sémantique* et des mesures de spécificité et de confiance ont été définies pour y remédier. Plus tard, d'autres ont utilisé le *clustering* pour trouver des classes de termes et en déduire des relations sémantiques, ainsi que des motifs comme ceux de Hearst, mais cette fois-ci pour réunir les termes qui participent à une conjonction dans un même cluster (par exemple, dans « X tels que Y_1 et Y_2 », où Y_1 et Y\down{2} sont cohyponymes de X).

Relations sémantiques et *deep learning*

Ce chapitre, qui est le principal ajout de la seconde édition de l'ouvrage, nous fait parcourir la quasi-totalité des techniques du *deep learning*, depuis les réseaux de neurones récurrents jusqu'aux transformeurs (BERT et compagnie), en passant par les réseaux convolutifs, les LSTM et biLSTM et le mécanisme d'attention. La présentation est bien structurée et ne présuppose qu'un minimum de connaissances dans le domaine. En plus, elle se place au juste milieu entre les présentations orientées code, que l'on trouve dans des ouvrages qui poussent comme des champignons depuis quelques années, et les présentations mathématiques avec des notations à vous faire dresser les cheveux sur la tête et que l'on trouve dans des ouvrages d'obédience plutôt mathématique.

Après deux sections courtes et purement introductives, la troisième section du chapitre s'intéresse à la modélisation des attributs à travers les plongements de mots, en commençant par la création de plongements à partir de textes. Il y est question d'abord des deux méthodes désormais classiques (*skip-gram* et sac de mots continu), de la possibilité d'utiliser des espaces métriques autres que l'espace euclidien pour les plongements et de l'introduction de données concernant le contexte des mots dans les plongements à travers les BiLSTM et les différents BERT. Ensuite, on s'intéresse à la création de plongements de mots (ou d'entités) à partir de structures ontologiques. *A priori*, cela est plus simple puisque tout est déjà structuré, mais on est aussi plus exigeant : plus qu'un simple contexte, on se propose de capter *toute* la structure de graphe de l'ontologie. Les problématiques afférentes sont discutées et accompagnées de références à de nombreux travaux dans le domaine.

Après avoir couvert toutes les facettes de la modélisation des attributs, on passe à la modélisation des relations sémantiques. Si, dans le paragraphe précédent on a parlé de « mots » et non pas de « nominaux », c'était pour faciliter l'approche par les réseaux de neurones. Mais ici on n'y échappe plus : une relation sémantique est forcément plus complexe qu'un « mot » et une première partie de cette section s'intéresse aux différentes manières de représenter des structures complexes (phrases, chemins dans un arbre syntaxique, chemins hiérarchiques dans WordNet, etc.) en tant que données d'entrée de réseau de neurones. Dans la

deuxième partie, il y a un véritable saut qualitatif : plutôt que d'essayer par tous les moyens de « linéariser » les structures complexes qui nous intéressent, on garde la structure de graphe et on utilise des réseaux de neurones de graphes, les travaux dans cette direction ne manquent pas à l'appel.

La section qui suit s'intéresse aux données. Après un court descriptif des corpus disponibles, on revient à la méthode, déjà décrite au quatrième chapitre, de supervision distante, méthode qui fournit des corpus très bruités. Pour remédier à ce problème de bruit, trois méthodes sont décrites : les réseaux adversariaux génératifs (où un générateur se bat contre un discriminateur pour séparer le grain de l'ivraie), les réseaux avec mécanisme d'attention et l'apprentissage à renforcement qui fonctionne comme un jeu, avec des états, des actions et des récompenses.

Arrivé·à ce point, on réalise que les sections 1 à 5 de ce chapitre n'ont fait que poser les fondements nécessaires à la section 6 qui est la plus longue et la plus dense du chapitre. De la même manière que l'on a considéré la modélisation de mots dans la section 3, il s'agit ici (enfin) de modélisation apprentissage de relations sémantiques. On considère tout à tour l'apprentissage de relations à partir de structures ontologiques, à partir de textes et à partir des deux réunis. Toutes les méthodes possibles et imaginables y passent, et rien que de les énumérer ferait sauter la limite de pages allouée aux notes de lecture !

Conclusion

Certains ouvrages de la collection *Synthesis Lectures on Human Language Technologies* sont focalisés sur des sujets assez pointus et, de ce fait, n'intéressent qu'un public restreint. Ce n'est pas le cas de celui-ci. Que ce soit par la culture débordante de ses auteurs ou par l'universalité de son sujet (on retrouve les relations sémantiques dans *tous* les domaines du TAL), cet ouvrage est un trésor d'informations et de connaissances et ne manquera pas d'enrichir son lectorat, qui peut être de tout niveau (du doctorant au chercheur confirmé). En plus, il se lit agréablement et fait preuve d'un bon équilibre entre érudition et technicité, entre grands principes et petits conseils pratiques. Nous le conseillons vivement et attendons avec impatience la troisième édition, qui sortira sans doute dans les années 2030-2035, et dans laquelle le *deep learning* sera relégué aux vestiges du passé, au profit de nouvelles technologies dont nous ignorons encore tout aujourd'hui.

Anne ABEILLÉ, Danièle GODARD. La Grande Grammaire du Français. *Éditions Actes Sud – Imprimerie nationale*. 2021. 2 628 pages. ISBN 978-2-330-14239-1.

Lu par **Dominique LEGALLOIS**

Université Paris 3 – Sorbonne nouvelle / LaTTiCe

Une année après le très bel ouvrage « Cette histoire de la phrase française » *(octobre 2020), les éditions Actes Sud font paraître la très attendue* Grande Grammaire du Français *dirigée*

par Anne Abeillé et Danièle Godard, avec la collaboration d'Annie Delaveau et Antoine Gautier. On rappellera que le projet a été initié en 2002 sous l'égide du CNRS et de la DGLFLF (délégation générale à la langue française et aux langues de France). Il a reçu le soutien de nombreux laboratoires de recherche et universités françaises ou étrangères. Une soixantaine de linguistes ont participé à la rédaction des deux volumes (plus de 2 500 pages) qui s'intègrent dans un coffret cartonné vert foncé. L'objet est en soi très beau. Il existe également une version Internet par abonnement, qui donne notamment la possibilité d'écouter des extraits sonores, car cette Grande Grammaire *accorde une place importante à la langue parlée.*

Encyclopédique, la *Grande Grammaire du Français* (désormais GGF) s'adresse à un grand public averti, mais aussi aux étudiants, enseignants et bien sûr aux linguistes. Son dispositif permet de multiples adresses : des zones grisées, souvent en début de section, présentent des éléments fondamentaux ; des descriptions plus fouillées ou plus techniques, à destination des lecteurs plus avertis, sont présentées en retrait. Des tableaux très utiles récapitulent des formes catégorisées, illustrées d'exemples. Cinquante fiches synthétiques sont proposées à la fin de l'ouvrage : elles décrivent la nature et la fonction de mots grammaticaux (*même, pour, où*, etc.) ou bien des fonctionnements grammaticaux (accord du verbe, de l'adjectif, inversion du sujet) ; leur utilité pour la préparation aux concours d'enseignement est évidente. Un tableau (en introduction) fait la synthèse de quelques-uns des changements terminologiques les plus importants, entre la terminologie de 1998 et celle de 2020. Un précieux glossaire de trente pages donne la définition des catégories employées. *Cinquante-six pages de références concluent l'ouvrage, mais la fin de chaque chapitre mentionne des repères bibliographiques en lien direct avec les questions traitées.*

L'une des caractéristiques principales de la GGF est qu'elle a été réalisée à partir des outils de la linguistique moderne : les corpus et les bases textuelles (par exemple, le Corpus de français parlé au Québec, les corpus Valibel, Frantext, le corpus « 88milSMS », Clapi, etc.) ont permis de calculer des fréquences définissant des usages (du standard et du non-standard) ou bien des variations (diatopiques, de registre) selon l'origine des données. En cela, la GGF se veut fondamentalement descriptive : non pas par souci de non-normativité, mais par souci de décrire la langue française telle qu'elle se parle et s'écrit, dans sa diversité même, depuis les années cinquante, dans une partie de la francophonie. En effet, les grammaires françaises « sur le marché », dont personne ne peut raisonnablement dire qu'elles sont normatives, se fondent encore sur l'écrit littéraire, mais surtout, en réalité, sur des exemples inventés appropriés. Si la GGF illustre également le plus souvent ses rubriques par des exemples de ce type (quatre mille exemples attestés jalonnent l'ouvrage), elle les invente en fonction des données empiriques recueillies et analysées. En cela, ces exemples reflètent l'usage et leur manipulation vient justifier un type d'analyse approprié. Ils sont de plus annotés, concernant leur acceptabilité, selon diverses catégories : inacceptable (*), inacceptable dans le contexte (#) (*Jean est arrivé, mais je ne sais pas où*), douteux (?), variable selon les locuteurs (%) (*il a mangé ici plusieurs personnes*), non conforme à la norme (*tu peux les faire manger la viande* – Louisiane) (!). La GGF n'est cependant pas une grammaire de l'usage, au sens strict du terme.

La GGF se compose de vingt chapitres, avec une structuration à la fois classique et particulière. Heureusement classique, car on ne voit pas comment une grammaire pourrait échapper aux catégories fondamentales de la phrase, du verbe, du syntagme nominal, etc. Particulière, parce que certains chapitres sont consacrés exclusivement à des catégories sémantiques (*la négation ; le temps, l'aspect, le mode*) et aussi parce que d'autres chapitres développent des notions que l'on s'attendrait à voir traitées dans une partie plus englobante, par exemple le chapitre « *La coordination et la juxtaposition* » reprend et complète celui sur « *La phrase* » (chapitre 1). « *Le verbe* » introduit beaucoup d'éléments repris et développés dans le chapitre suivant, « *Les constructions verbales fusionnées* ». Le titre de ce chapitre atteste, par ailleurs, certaines nouveautés terminologiques (plus habituellement on parlerait de *complexe verbal* plutôt que de *fusion*). Celui intitulé « *Les proformes* » traite de la présentation d'une catégorie plus générale que celle traditionnelle des pronoms (ainsi y figurent également toute forme dont l'interprétation nécessite le contexte ou la situation : les déterminants démonstratifs et possessifs, les adverbes interrogatifs (*quand*), les adjectifs *tel* et *quel*, etc.). Contrairement aux autres types de subordonnées, la subordonnée complétive ne fait pas l'objet d'un chapitre – non plus d'ailleurs que la subordonnée participiale, quelque peu ressuscitée sans trop que l'on sache pourquoi. Les quatre derniers chapitres sont des chapitres d'interface particulièrement bien documentés (cf. la forme sonore des énoncés, l'ancrage des énoncés dans l'énonciation, les écritures numériques…).

Parmi les choix innovants de la GGF, on note par exemple la notion de « marqueur », qui vient préciser la nature prépositionnelle de *de* et *à* employés devant l'infinitif (*Paul continue de / à travailler*) et qui se justifie par le fait que la préposition peut être séparée de l'infinitif par *ne pas*, *ne plus* – elle est alors considérée comme introducteur de syntagme verbal (comme le subordonnant (conjonction) *que* est marqueur, car il introduit un syntagme phrastique). *Ici*, *là*, *là-bas*, *dehors*, *partout*, etc. traditionnellement considérés comme des adverbes, sont, dans la GGF, classés parmi les prépositions sans complément en raison de la difficulté à les employer entre l'auxiliaire et le participe passé (ou devant l'infinitif) : **Paul est dehors allé* ; **je ne veux pas ici aller*. Pour la même raison, *où* perd son statut d'adverbe et même de relatif « plein » : il est une préposition soit interrogative, soit relative (et relative sans antécédent), soit concessive (*où que tu ailles*). Remarquons tout de même que *où* est donné comme adverbe de lieu dans le chapitre consacré aux proformes : cette petite incohérence montre la difficulté qu'ont dû rencontrer les directrices de la GGF pour homogénéiser chez les soixante contributeurs, la terminologie et les analyses. La fonction *coordonné* s'applique à tous les membres d'une coordination ; la fonction *extrait* aux membres disloqués ou aux thèmes suspendus.

Autres spécificités : *certains* (dans *certains m'ont fasciné*) ou encore *les miens*, *les tiens*, etc. sont analysés comme des syntagmes nominaux sans nom, et non comme des pronoms. Ainsi *certains* garde dans l'exemple sa nature de déterminant devant un nom en ellipse. Il s'agit là justement de proformes. En revanche, *ce dernier* est bien un pronom, car il ne peut être modifié. La notion d'*intransitif* est étendue aux verbes à complément oblique (*j'ai rêvé de vous*) : l'impotent transitif

indirect est jeté aux orties – qu'il y reste ! On parle dans la GGF d'*omission du complément* plutôt que d'emploi absolu du verbe. Les notions de voix ou de diathèse sont rangées dans un vieux tiroir : *alternances de valence* suffira. Le complément direct perd son objet, ce qui permet d'éviter la distinction entre COD et complément essentiel. *Phrase désidérative* remplace *phrase à l'impératif*. Le terme *proposition*, en fait assez peu employé, est restreint à un emploi particulier : il désigne le contenu sémantique d'une phrase déclarative et non plus une configuration formelle.

On peut bien sûr faire quelques critiques, nécessairement éparses ici. C'est d'ailleurs le propre d'une grammaire que de se soumettre à la question du linguiste. Ainsi, la GGF entérine la fonction d'*ajout* à la place de l'encombrant *complément circonstanciel* ; cependant, la catégorie de l'ajout concerne des cas qui nous semblent relativement différents puisque cette fonction s'applique non seulement aux éléments circonstants, mais également aux adjectifs considérés parfois en emploi attribut (*elle est partie furieuse*), aux incises (*Lou, je crois, a terminé son travail*), aux pronoms contrastifs (*Paul viendra, lui*) ou quantifieurs (*les élèves viendront tous*). La fonction *ajout* est donc très générale, ce que reconnaissent explicitement les autrices. De même, on regrette aussi que la fonction *oblique* (plus universelle que celle de *complément indirect*) soit trop englobante : elle ne permet pas d'établir de distinction entre certains compléments indirects et les compléments datifs (*Paul parle de Marie à Marc*), alors que les pronominalisations de ces compléments sont pourtant bien différentes (cette remarque vaut en fait pour presque toutes les grammaires du français). On peut aussi se demander si la multiplication des fonctions n'ajoute pas une certaine complexité descriptive : si le linguiste peut y trouver son compte, le pédagogue, lui, a du souci à se faire… Sauf erreur de notre part (car la GGF est un océan vaste à naviguer), la particularité sémantique et syntaxique des verbes labiles (*Paul cuit les pâtes* vs *les pâtes cuisent*), qui aurait pu illustrer l'alternance de valence, ne fait pas l'objet d'une description, ni même d'une mention, alors que ces verbes (verbes ergatifs selon certaines terminologies) représentent un nombre important de lexèmes verbaux parmi les plus courants. Ces quelques remarques – qui pourraient, elles aussi, être discutées – peuvent se justifier globalement par la difficulté à embrasser et apprécier en peu de temps, un volume tout à fait considérable d'aspects descriptifs nouveaux : ce qui est fondamentalement présent dans la GGF est une logique descriptive et analytique dont l'appropriation ne peut être que progressive.

En résumé : la GGF est une somme considérable qui fait déjà date dans l'histoire de la grammaire du français. C'est un ouvrage de référence dont l'aspect encyclopédique et l'approche « usage » viennent combler une lacune dans la description du français. Il s'agit là d'un magnifique outil dont on n'a pas fini de manipuler les pages.