Exploration informatisée de l'écriture de Gonzalo Fernandez de Oviedo y Valdes dans la *Historia General y Natural de las Indias*

Henri P. DE STAMPA

Abstract. At the beginning of the sixteenth century, Gonzalo Fernandez de Oviedo y Valdes decides to write *la Historica General y Natural de Las Indias*. Official chronicler of the spanish conquest of America, he describes expeditions and historical events, he shows natural and social universe.

The exploration of this work of Oviedo within the framework of research of the ARIT (Laboratoire d'Analyse Relationnelle Informatique des Textes, Université de Paris VIII) is a description of the terminology when the chronicler uses the three most frequent indian words of the natural elements vocabulary, that is to say *huracán*, *sabana*, and *jagüey*.

The corpus groups together all sentences including, at least, one of these indian terms. Each component of the sentence (grammatical or lexical words and punctuation marks) is identified in the linear order with its morphological, functional and spacial characteristics.

The data file necessary for such a compilation contains 150,746 informations. After several experiences, three kinds of data processing tools are chosen: *Lotus 1-2-3*, *dBase III Plus* and *Spad-N* which is a factor analysis system.

The research links together two main phases.

- During the first one, we try to have a good knowledge of sentences including indian words. We discover the subjects of the corpus in connection with such words, environment, conquerors' action, meaning of new terms, importance, modalités and complexity of phrases according to a specific typology.
- In the second phase, the use of a factor analysis system reveals a lot of invisible phenomenons for human readers when we distinguish the different types of discourse. Its results are instructive and very varied: 250 relations are significant. They mainly concern morpho-syntactic characteristics and the position of verbal nucleus in phrases of which length may reach 195 elements.

Such a research finds its efficiency in linguistic and organizational principles: at any moment each component of a sentence may be observed in the most exhaustive context, data processing systems have to offer a large variety and an excellent complementarity of functionalities; actually, a flexible, evolutive and complete organization permits investigations which were not conceivable at the beginning of the work.

✉ Rue Belgrand, 70; F–75020 Paris (France).
Fax : +33 1 40 49 14 52

Keywords: Linguistic, spanish, factor analysis system.

Mots-clés : Linguistique, espagnol, analyse factorielle.

Dans les années qui précédèrent le cinquième centenaire de la découverte de l'Amérique plusieurs membres du Laboratoire d'analyse relationnelle informatique des textes à l'Université de Paris VIII consacrèrent leurs travaux aux principaux documents descriptifs de la conquête. Des recherches significatives furent en particulier conduites sur l'introduction de terminologies amérindiennes et c'est ainsi que furent surtout relevées du point de vue de leur appartenance à diverses langues, les indo-américanismes dans *La Historia General y Natural de las Indias* de Gonzalo Fernandez de Oviedo y Valdes.

Les travaux que nous nous proposons d'évoquer aujourd'hui font très largement appel à l'informatique mais, faute d'un enregistrement magnétique intégral de *La Historia General y Natural de las Indias*, ils ont pour point de départ certains relevés manuels et classements proposés par Roger VALENSI[1].

L'étude est consacrée à une vaste chronique (cinquante livres, des centaines de pages) écrite au début du XVI^{e} siècle. Elle a été précédée quelques années plus tôt par un sommaire détaillé qui en constitue le résumé. Son auteur, Oviedo, est à la fois chroniqueur et acteur. On le présente souvent comme un autodidacte. C'est assurément un érudit et un homme d'action. Il a surtout vécu en Espagne et en Italie du Nord pendant la période glorieuse des Sforza. Il a été capitaine et administrateur territorial. Il a personnellement combattu et souffert pendant la conquête, ce qui l'autorise à parler en connaisseur des expéditions et des évènements historiques. Il présente sa vision de l'univers naturel et social, il décrit la faune, la flore et les modes de vie.

Oviedo s'exprime généralement bien. Ses phrases sont parfois longues (jusqu'à 200 mots); quelques-unes sont maladroites, comme s'il avait dicté rapidement à un secrétaire.

L'exploration informatisée de l'écriture d'Oviedo dans cette œuvre est une description terminologique du discours dans lequel s'intègrent les trois indo-américanismes les plus fréquents du lexique des éléments naturels, à savoir *jagüey*, sorte de puits peu profond creusé bien souvent à la main pour étancher sa soif, *huracán*, l'ouragan et *sabana*, la savane.

Le corpus de travail rassemble les énoncés qui contiennent l'un au moins de ces indo-américanismes. L'ensemble des énoncés comprenant un même indo-américanisme constitue un sous-corpus.

---

[1] Thèse inédite, Université de Paris VIII, 1986.

Notion fondamentale pour cette démarche, l'énoncé est une succession de formes, c'est-à-dire de composants significatifs ou structurants : les mots grammaticaux ou lexicaux et la ponctuation, repérables dans l'ordre linéaire. Il est délimité par le point qui achève l'énoncé précédent et par celui qui l'achève à son tour.

Tout au long de l'étude, la forme, en tant qu'élément constitutif de l'énoncé, peut être étudiée dans l'exhaustivité de son contexte textuel : elle se voit toujours associer l'ensemble de ses caractérisques morphologiques, fonctionnelles et spatiales.

Les notions traditionnelles de proposition et de groupe n'ont pas été retenues. À l'expérience elles sont en effet apparues comme des sous-ensembles de l'énoncé tellement fragmentés et imbriqués les uns dans les autres qu'il n'était pas toujours possible de les exploiter rationnellement par des moyens informatiques.

Choisir trois indo-américanismes parmi plusieurs dizaines revient à limiter délibérément le champ de l'étude mais nous nous situons bien cependant dans une problématique de grand corpus avec quelque 8 000 formes.

La base de données constituée pour l'exploration informatisée de l'écriture d'Oviedo se présente comme un vaste tableau contenant 150 746 informations. Les outils techniques de compilation et d'analyse de données retenus au moment où l'étude a été entreprise, il y a 9 ans, consistent en trois types de progiciels qui présentent des fonctionnalités très différentes et qui doivent s'utiliser dans l'ordre suivant.

En premier lieu, le tableur *Lotus 1-2-3*. Il sert à enregistrer les formes du corpus et l'ensemble de leurs caractéristiques dans des tableaux de grande dimension. Il a pour inconvénient la multiplication des redondances mais deux avantages majeurs :

- la visualisation instantanée et exhaustive des données enregistrées,
- la simplicité de la correction et de l'édition sur papier.

Deuxièmement un système de gestion de base de données. La particularité d'un tel système est qu'un tri, même maladroit, n'altère jamais l'ordonnancement initial des informations. Plusieurs progiciels essayés ont présenté des insuffisances par rapport à nos besoins et nous avons finalement retenu *dBase III Plus*. Le transfert des données du tableur au système de gestion de bases de données a été réalisé par des procédures de traitement automatique.

Troisièmement un dispositif d'analyse factorielle. Nous avons d'abord expérimenté *Tri-deux*, ensemble de logiciels conçus et offerts par Philippe

CIBOIS[2] mais finalement notre choix s'est arrêté, après juin 1991, sur *Spad-N*, Système Portable pour l'Analyse des Données, diffusé par le C.I.S.I.A.[3] *Spad-N* est en effet très convivial pour un utilisateur non professionnel et présente des fonctions statistiques qui complémentent utilement l'analyse factorielle proprement dite.

L'exploration comporte deux grandes étapes.

— Dans un premier temps, nous faisons connaissance avec le contenu et la forme de chaque sous-corpus. Nous découvrons les sujets traités, parmi lesquels l'environnement, l'action des conquérants, la définition des termes nouveaux. Nous nous faisons une idée de l'importance relative, des modalités et de la complexité des énoncés composant chaque sous-corpus. L'examen successif de toutes les particularités syntaxiques et sémantiques permet de différencier les sous-corpus. Si l'on s'en tient aux résultats les plus pertinents, après avoir distingué des familles d'énoncés principalement définitoires, secondairement définitoires et non définitoires des trois indo-américanismes, il apparaît que :
  - pour tous les sous-corpus, les énoncés principalement définitoires sont les moins complexes et les moins longs,
  - les énoncés secondairement définitoires sont les plus complexes et les plus longs quand il est question de *huracán* et de *sabana*,
  - les énoncés non définitoires sont compliqués et très longs quand il s'agit de *sabana*,
  - enfin, *huracán* est le sous-corpus le plus définitoire.

  Par ailleurs, pour 64 caractéristiques propres aux formes, il est constitué un relevé des fréquences les plus élevées attribuables à chaque sous-corpus, ce qui lui confère toute sa spécificité à ce stade de l'étude :
  - le sous-corpus *jagüey* a 19 caractéristiques dont 2 d'ordre sémantique,
  - *huracán* a 29 caractéristiques dont 5 d'ordre sémantique,
  - *sabana* a 16 caractéristiques dont 5 d'ordre sémantique.

  Voici des exemples de particularités :
  - Avec *jagüey* Oviedo use plus particulièrement du conditionnel, des pronoms, des substantifs sujets, des compléments directs et des circonstants

[2] LISH, Laboratoire Informatique pour les Sciences de L'Homme, 54, boulevard Raspail, F–75006 Paris (France).

[3] Centre International de Statistiques et d'Informatique Appliquées, 1, avenue Herbillon, F–94160 Saint-Mandé (France).

notionnels. L'emploi des noms propres de lieu est très soutenu ainsi que celui des termes techniques ou décrivant les produits de l'activité humaine.
- Avec *huracán* se démarquent le premier rang des formes verbales, la première personne, l'usage du parfait, l'adjectivation et la complémentation du nom, les circonstants de temps et de quantification, le recours à des champs sémantiques très particuliers : quantification, classement, jugement, chronologie, univers urbain et géographique, domaine de la religion. L'auteur évoque ses actions personnelles et donne son avis.
- Dans *sabana* dominent les compléments spécifiques de lieu, les circonstants de lieu, les termes de localisation, les repères sociologiques, les noms propres de personne. L'auteur use du présent.

— Au cours de la deuxième étape, nous cherchons à savoir comment s'exprime Oviedo quand, pour tel des indo-américanismes, il utilise un certain type de discours. Le corpus comprend en effet des indo-américanismes dans du récit, dans des titres et dans des passages où le narrateur s'implique personnellement. Cette deuxième phase a recours au dispositif d'analyse factorielle et tout particulièrement à sa fonctionnalité d'analyse des correspondances.

11 grands thèmes sont soumis à l'analyse :
- les modalités d'utilisation du verbe (mode, temps, personne, rang des formes verbales)
- le repérage spatial des noyaux verbaux (identification de bases primaires, secondaires ou tertiaires)
- les formes verbales,
- la fonction syntaxique des substantifs,
- la fonction syntaxique des pronoms,
- les déterminants,
- l'adjectivation non médiate,
- le complément du nom,
- les sujets explicites,
- les compléments du verbe,
- les circonstants.

La productivité de cette approche est inégale selon les sujets abordés mais ses résultats discriminants sont instructifs et d'une grande diversité : 250 correspondances sont considérées comme pertinentes; elles concernent

principalement les caractéristiques morphosyntaxiques et le positionnement du noyau verbal dans l'énoncé.

— Dans le récit les fréquences élevées sont réparties à peu près à égalité entre les trois sous-corpus mais, bien entendu, elles les caractérisent différemment dans la mesure où, par convention, une même fréquence élevée ne peut être attribuée qu'à un seul sous-corpus.
— Dans les passages où le narrateur s'implique, *huracán* rassemble 43 % des correspondances, contre seulement 38 % pour *sabana* et 19 % pour *jagüey*.
— Les titres présents dans le corpus concernent tous *huracán*.

Ceci révèle sans ambiguïté une différenciation significative du discours d'Oviedo quand il veut traduire le phénomène incommensurable de l'ouragan, bien plus effrayant, semble-t-il, que la peur de manquer d'eau potable généralement présente avec *jagüey*.

La base de données qui a été construite et les modalités d'exploitation qui ont été proposées constituent une source d'information intéressante sur la diversification de l'écriture d'Oviedo, notamment en fonction de l'utilisation des indo-américanismes, des noyaux verbaux et du type de discours.

L'efficacité de la recherche tient en grande partie à l'enchaînement judicieux d'un tableur, d'un système de gestion de bases de données et d'un dispositif d'analyse factorielle, enchaînement qui permet de disposer tout au long des travaux d'une grande variété et d'une excellente complémentarité des fonctionnalités. Grâce à cette organisation souple, évolutive et complète le chercheur ne risque à aucun moment d'être prisonnier d'un cahier des charges contraignant qui deviendrait une gêne quand se présenterait l'opportunité d'une investigation non imaginable à l'origine des travaux.

Cet outil permet de conserver en permanence la forme dans son contexte minimal que constitue l'énoncé, en lui associant chaque fois les trois informations linguistiques fondamentales pour nous dans cette étude :
– l'ordre linéaire,
– la classe morphologique,
– les fonctions syntaxiques.

Ces travaux se situent dans une perspective ouverte; ils pourraient servir de base à des études pouvant bénéficier des évolutions actuellement prévisibles de l'outil informatique.

Bibliographie

1. Ouvrages

HUYNH-ARMANET (Véronique) : 1976, *Recherches sur la structuration syntaxique de l'espagnol contemporain* (Lille III et Honoré Champion).

HUYNH-ARMANET (Véronique) : 1977, *Des structures syntaxiques de l'espagnol à l'analyse relationnelle des textes* (Paris : Honoré Champion).

POTTIER (Bernard) : 1972, *Introduction à l'étude linguistique de l'espagnol* (Paris : Ediciones Hispano-Americanas).

HAGÈGE (Claude) : 1986, *L'homme de parole* (Paris : Fayard).

2. Articles

HUYNH ARMANET (Véronique) : 1977, *L'analyse textuelle du langage, l'infra-littérature en Espagne aux XIX^e^ et XX^e^ siècles* (Grenoble : Presses Universitaires de Grenoble).

SANCHEZ PEREZ (Francisco-Javier) : 1990, *Les logiciels intégrés et plurilinguistiques en description contextuelle* (Paris : Cahiers de linguistique relationnelle informatique, Laboratoire ARIT, Université de Paris VIII).

Annexe

Les tableaux 1 et 2 sont des fragments extraits de la base de données constituée sur tableur.

Tableau 1

Exemple de repérage des formes dans l'espace textuel

Tome	Livre	Chapitre	Page	Côté	Paragr.	Énoncé	nuF	nuA	Forme
1	6	3	146	b	3	3	1	s	*Huracán*
1	6	3	146	b	3	3	2	s	,
1	6	3	146	b	3	3	3	s	*en*
1	6	3	146	b	3	3	4	s	*lengua*
1	6	3	146	b	3	3	5	a	*desta*
1	6	3	146	b	3	3	6	b	*desta*
1	6	3	146	b	3	3	7	s	*isla*
1	6	3	146	b	3	3	8	s	,
1	6	3	146	b	3	3	9	s	*quiere*
1	6	3	146	b	3	3	10	s	*decir*
1	6	3	146	b	3	3	11	s	*propriamente*
1	6	3	146	b	3	3	12	s	*tormenta*
1	6	3	146	b	3	3	13	s	*o*
1	6	3	146	b	3	3	14	s	*tempestad*
1	6	3	146	b	3	3	15	s	*muy*
1	6	3	146	b	3	3	16	s	*excesiva*
1	6	3	146	b	3	3	17	s	;

La forme est inscrite dans la base de données en colonne 10 avec son repérage spatial et l'ensemble de ses caractéristiques. Une forme qui résulte de la concaténation de plusieurs autres figure plusieurs fois et fait l'objet d'une analyse pour chacun de ses composants.

Les huit premières colonnes donnent le repérage dans le texte : tome, livre, chapitre, page, côté de page, paragraphe, numéro d'énoncé dans la page, numéro de la forme dans l'énoncé. La neuvième colonne identifie le type d'analyse : formes simples ou concaténées.

Tableau 2

Exemple d'analyse des formes du corpus

Forme	Cl. morph.	Cl. fonc.	Fonct. synt.	Temps ou genre	Mode ou champ sémantique	Rang ou Nombre	Personne	Temporalité ou Glissement	Discours
Huracán	SUB	Z	S	M	N	S	Z	AME	R
,	PON	INT	Z	Z	Z	Z	Z	Z	R
en	PRP	Z	RS	Z	L	Z	Z	Z	R
lengua	SUB	Z	CN	F	S	S	Z	Z	R
desta	PRP	Z	RS	Z	Z	Z	Z	Z	R
desta	ADJ	D3	Z	F	Z	S	Z	Z	R
isla	SUB	Z	A2	F	N	S	Z	Z	R
,	PON	INT	Z	Z	Z	Z	Z	Z	R
quiere	VP	Z	BP	PR	IND	3	3	Z	R
decir	VP	Z	P0	PR	INF	Z	Z	Z	R
propriamente	ADV	Z	CN	Z	Z	Z	Z	Z	R
tormenta	SUB	Z	P0	F	N	S	Z	Z	R
o	CON	Z	RC	Z	Z	Z	Z	Z	R
tempestad	SUB	Z	P0	F	N	S	Z	Z	R
muy	ADV	Z	CQ	Z	Q	Z	Z	Z	R
excesiva	ADJ	Z	A0	F	A	S	Z	Z	R
;	PON	INT	Z	Z	Z	Z	Z	Z	R

La forme est inscrite en première colonne. Les colonnes 2 à 10 enregistrent ses caractéristiques : classe morphologique, classe fonctionnelle, fonction syntaxique, temps ou genre, mode ou champ sémantique, rang de forme verbale ou nombre, personne de forme verbale, temporalité ou glissement sémantique, type de discours dans lequel elle est utilisée.