

ANNOTATION DE CORPUS LINGUISTIQUE ET TEI

Remi JOLIVET
Université de Lausanne (Suisse)

I am a poor lonesome linguist...

Il ne s'agit pas ici d'encenser ou de vomir la TEI (Text Encoding Initiative). Mais, au terme d'une tentative d'application concrète et effectivement largement aboutie¹, en tout cas très satisfaisante à mes yeux, de se demander si la démarche a néanmoins sa place dans la pratique du "linguiste ordinaire" travaillant dans les conditions usuelles, c'est-à-dire en solitaire ou presque, en tout cas pas au sein de grandes équipes nationales ou internationales, multidisciplinaires et bien dotées en ressources humaines et budgétaires.

Le corpus

Comme il n'y a pas de définition généralement admise de ce qu'est un corpus linguistique² le mieux est de partir de l'observation de ce que des linguistes considèrent comme "corpus". Il peut s'agir d'une simple liste: liste de monèmes; de paires minimales; de contextes d'apparition d'un élément (concordances) etc. Ou bien des données obtenues via un questionnaire³. On s'en tiendra ici à une forme usuelle et fondamentale de corpus: un ensemble d'énoncés produits par un ou plusieurs locuteurs, dans la langue sur laquelle le linguiste travaille.

Cela peut se résumer à la fixation graphique (transcription) d'événements langagiers, avec quelques informations complémentaires. Par exemple cet extrait du corpus de Denise François⁴:

		I,27-30	773
	LS	α me se lkalkyl ã	
	LSe	me ãfã jãna kelkəzyn {	ki dwəv pase
	DMe		mm
	LSe	swadizã sã... {	egzamã
I, 28	LS		α ba fəsemã { le pəmjek
	DMe		{ wi kãtəna la mwəjen
	LSe	{ sə kə la mwəjen	
	LS	{ ja la mwəjen aləv se nəvmal {	se nəvmal ã
	LSe		{ epi aləv kəm el e pə mal pəvə... kwə
		ã pti pə	
	LS	wi	
	LSe	aləv sa ...s ...se {	sa zu kãmem ã pti kəl ladsy { ã
	LS		{ wi αα { wi α me kãmem
		pəv... fo kənet kə dã lkalkyl ã (claquement de langue)	
	LSe	α wi, sa wi	
	LS	ty se... {	
	DMe		{ setadiə ksa zu ã kəl avã me ty se apə finalmã... ɔne
		pəi dã lə... ãfã syə plas ɔne pəi dã lbã	
	LS	α wi me fo lsezis {	aləv si elã ã pəblem kə, kel aiv pə
	DMe		{ wi
	LS	al definis {	
	DMe		{ wi ɔ
I, 29	LS	ã se bo {	se... {
	LSe		{ kəmæk la, el a fe ãn egzamã, elmadi
		{ aləv el mə di sez əpekəsjə ɔ ete bən, yn pasti	
	LS		{ wi
	LSe		{ dy pəblem, ãfã el se pə ãkəv ã, yn pasti dy pəblem, el a fe
		ynə, yn dikte kete pə pəv tko məvez aləv ɔ... si se lãsãblə ki	
		kəv {	
	DMe		{ wi wi se sa
	LSe		{ e pə aivə a... avvək le pəv... pə... vuly, mætnã... (rire) e pə avvək lminiməm, ã, me
	LS		{ ɔ vəkə, ɔ vəkə... (rire) tu sa ã
	DMe		{ sa ja yn pəv də fəkm e tu sa ki kəv aləv...
	LS		{ mm
	LSe		{ e pi aləv mætnã jãna ki diz ki
	LS		{ e pi... dkapasite

Figure 1. Corpus de Denise François

Mais on peut enrichir la transcription de l'enregistrement sonore par des annotations linguistiques. Ainsi dans ce corpus d'oubikh préparé par Georges Dumézil⁵:

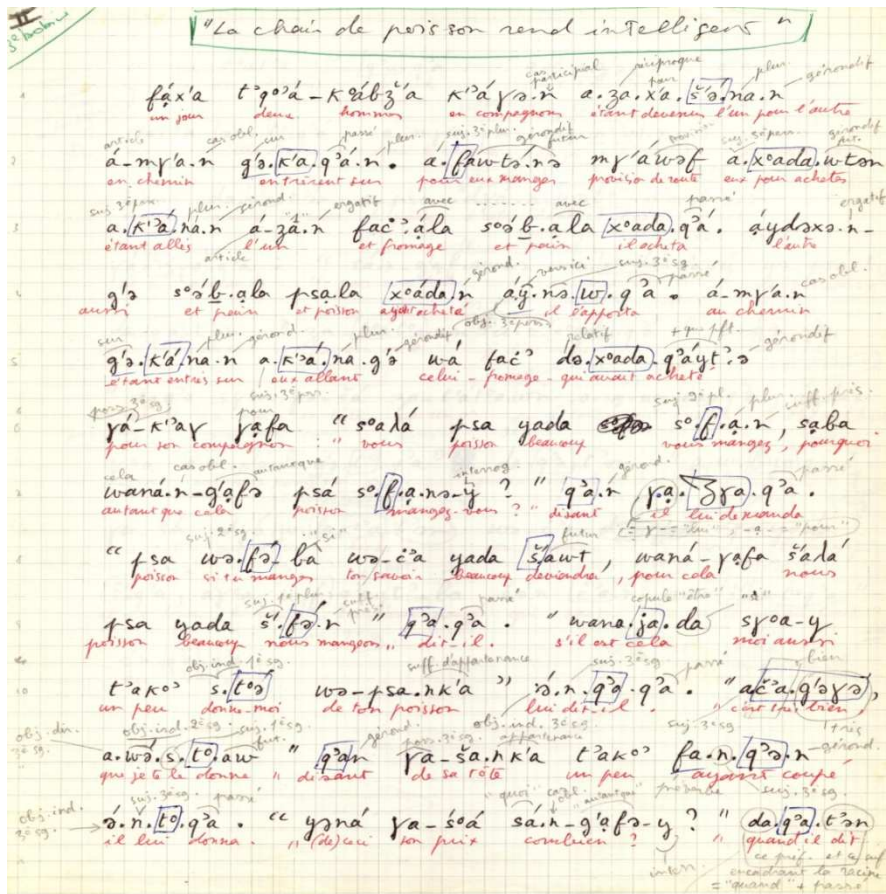


Figure 2. Corpus de Georges Dumézil (source: LACITO)

Aux éléments de la transcription de l'enregistrement sonore est associée une traduction des syntagmes. Des gloses identifient spécifiquement les éléments grammaticaux. La transcription et ses enrichissements se présentent en lignes parallèles, mettant en correspondance les éléments des différents niveaux d'information en les alignant.

La création de ces corpus suppose des compétences multiples (notation phonétique; analyse grammaticale) et s'appuie sur des conventions partagées (tableau phonétique; terminologie).

Ainsi constitués et présentés, ils sont immédiatement lisibles par les linguistes qui les utilisent. Cette utilisation, dans la pratique descriptiviste, présente généralement deux aspects. Le corpus est d'abord la source d'information du linguiste, qui l'analyse pour dégager les régularités constitutives de la structure de la langue. Ensuite, lorsqu'il s'agit de présenter les résultats de cette analyse⁶, le corpus servira de répertoire d'exemples ou d'illustration à l'usage du lecteur.

L'apparition des micro-ordinateurs a grandement facilité la correction, l'enrichissement, la présentation ou la diffusion de tels corpus. Il devient possible de les associer à des enregistrements, audio ou vidéo. La figure suivante présente une fenêtre de l'un des nombreux logiciels - ici il s'agit d'ELAN⁷ - développés pour faciliter le travail du linguiste. En bas de cette fenêtre on retrouve les lignes parallèles caractéristiques de la forme de ce genre de corpus.

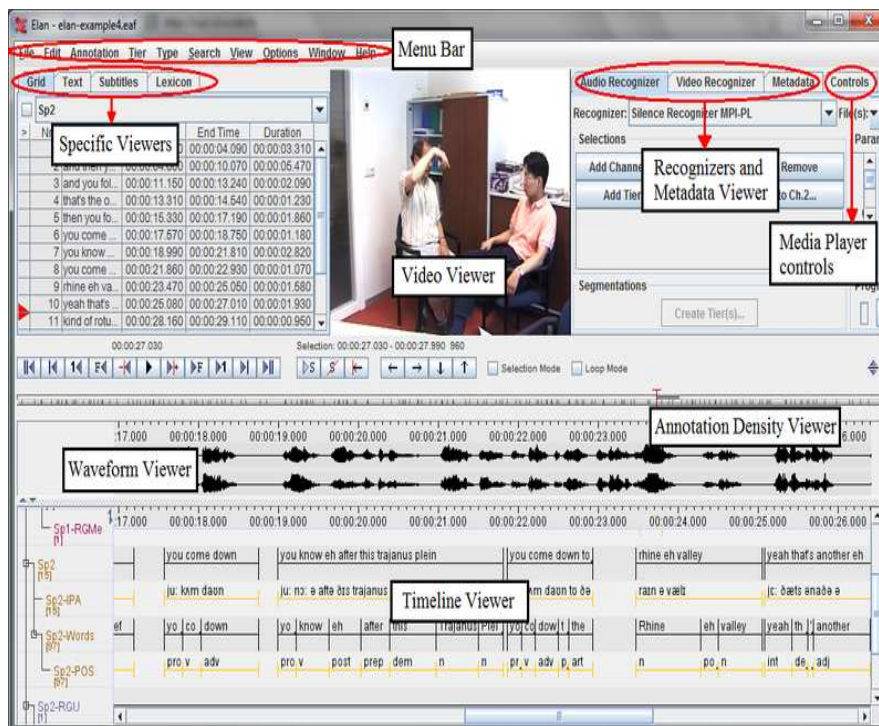


Figure 3. Une fenêtre d'ELAN

Et, finalement, l'exploitation - la recherche de phénomènes pertinents par rapport à une question donnée - est accélérée grâce à la vitesse de "lecture" de la machine et aux facilités qu'elle offre pour la recherche de chaînes de caractères. Cependant, pour le linguiste, l'apparence ne change guère. Voici comment s'affiche le corpus Oubikh de Dumézil dans la collection Pangloss du LACITO⁸, site web archivant de nombreux corpus linguistiques:

Accueil > Accueil Pangloss

Eating fish makes you clever[Ⓢ]
Langue : Ubykh(uby)

Chercheur(s) : Dumézil, Georges Locuteur(s) : Saniç, Tevfk

1:36

Lecture en continu :

Transcription par phrase Transcription du texte complet
 Phonologique Mot à mot

Traduction par phrase Traduction du texte complet
 FR EN FR EN

S1 fɛx'a t'q'á-k'ábɟ'a k'áɣə.n aza.x'a.š'ə.na.n á-my'a.n g'ə.k'a.q'á.n.
 fɛx'a t'q'á-k'ábɟ'a k'áɣə.n aza.x'a.š'ə.na.n á-my'a.n g'ə.k'a.q'á.n
 un jour deux hommes en compagnons étant devenus l'un pour l'autre en chemin entrèrent sur
 Un jour, deux hommes prennent la route ensemble.

S2 a.fawtə.nə my'áwəf a.x'ada.wtən a.k'á.na.n á-za.n fac'ála s'əb.ála x'ada.q'á.
 a.fawtə.nə my'áwəf a.x'ada.wtən a.k'á.na.n á-za.n fac'ála s'əb.ála x'ada.q'á
 pour eux manger provision de route eux pour acheter étant allés l'un et fromage et pain il acheta
 Ils vont s'acheter des provisions pour la route. L'un achète du fromage et du pain,

S3 áydxə.n-g'ə s'əb.ála psə.la x'áda.n á.y.nə.w.q'á.
 áydxə.n-g'ə s'əb.ála psə.la x'áda.n á.y.nə.w.q'á
 l'autre-aussi et pain et poisson ayant acheté il apporta
 l'autre du poisson et du pain.

Figure 4. Corpus de G. Dumézil dans la collection Pangloss du Lacito

L'informatique multiplie donc les possibilités du linguiste mais sans modifier fondamentalement la forme des objets sur lesquels il travaille. Bien sûr, puisqu'il s'agit d'informatique, il y a une - et même plusieurs - représentations plus proprement informatiques de ces objets. Elles sont destinées au fonctionnement de l'ordinateur et ne concernent le sujet humain que s'il est informaticien. Le linguiste n'est pas concerné, même si certaines connaissances dans ce domaine ne peuvent pas nuire, comme des connaissances en mécanique ne peuvent pas nuire à l'automobiliste.

La TEI - TextEncoding Initiative

Le projet démarre à la fin du XXème siècle, en 1987. Son noyau est constitué de Recommandations (Guidelines for ElectronicTextEncoding and Interchange⁹) qui fixent des conventions de représentation du texte et des informations qu'on peut vouloir lui associer:

- métadonnées,
- structuration (en chapitres, en paragraphes, en actes, en répliques, en strophes, en vers etc.),
- éléments d'analyse de tous niveaux (marquage - "balisage" - des "entités nommées": noms propres, dates; des tours de parole; d'une analyse syntaxique etc.),
- informations relatives aux principes mêmes de ces analyses,
- informations relatives à l'"histoire" du document électronique,
- etc.

Pour permettre cette représentation, un ensemble copieux d'éléments (env. 550) - dont l'ambition est de couvrir tout ce que l'on peut associer d'information à un texte - a été progressivement défini. Ces éléments peuvent être spécifiés par l'association de propriétés (attributs) et des règles explicites président à leurs combinaisons hiérarchiques. Ces inventaires d'éléments et d'attributs sont ouverts, car la TEI est adaptable, vivante et aussi pérenne et stable, qualités remarquables quand on sait que bien des grands projets informatiques n'aboutissent pas, ou au prix d'une réduction drastique des ambitions initiales¹⁰.

Eléments et attributs sont actuellement exprimés en suivant les règles d'un langage de représentation informatique qui a aujourd'hui de multiples applications (de multiples "dialectes"): XML pour 'eXtensible Markup Language', langage de balisage extensible.

La TEI, comme démarche et comme ensemble de conventions, a

d'évidentes qualités:

- l'idée même de standardisation pour l'échange des textes électroniques,
- mais une standardisation qui ne constitue pas un carcan et reste souple et adaptable à des besoins spécifiques,
- la richesse des domaines (types de textes) couverts,
- mais une modularité qui permet de ne pas s'encombrer de ce qui est inutile dans le cadre d'une application particulière,
- une grande simplicité (clarté) conceptuelle,
- une communauté très réactive (et francophone¹¹),
- une documentation très riche,
- une mise à disposition d'outils pour la définition d'un schéma d'encodage TEI et le contrôle de sa conformité,
- un logiciel, commercial mais au coût très raisonnable pour le personnel académique, qui facilite l'encodage et le traitement des documents XML¹².

Mais elle présente aussi des défauts¹³ ou des inadaptations au travail du "linguiste ordinaire":

- la compréhension et la maîtrise des conventions exigent un effort certain et un investissement en temps non négligeable. La dernière version des Recommandations (Guidelines) compte plus de 1600 pages, en anglais. Et il est risqué de les survoler trop rapidement.
- la souplesse du système et son adaptabilité sont des qualités qui ont leur revers. La possibilité d'encoder les mêmes phénomènes (une analyse syntaxique par exemple) de multiples façons, parfois fort différentes, peut être source d'hésitation ou de confusion.
- plus il est riche, plus il est détaillé - et donc potentiellement utile - et plus l'encodage est long¹⁴ et exposé à des erreurs. Quant au résultat il est rapidement peu lisible pour un lecteur humain, comme le

montre l'encodage de la phrase¹⁵:

"It was also a crucial year for me because on June 18, 1954, I began serving a sentence in state prison for possession of marijuana"

```
<s>
<cl type="finite-declarative" function="independent">
<phr type="NP" function="subject">It</phr>
<phr type="VP" function="predicate">
<phr type="V" function="verb-main">was</phr>
also
<phr type="NP" function="predicate-nom.">a crucial year for me</phr>
</phr>
<cl type="declarative-finite" function="dependent-causative">because
<phr type="PP" function="sentence_adverb">on June 18, 1954</phr>,
<phr type="NP" function="subject">I</phr>
<phr type="VP" function="predicate">
<phr type="V" function="verb-main">began serving</phr>
<phr type="NP" function="complement">a sentence in state prison
<phr type="PP" function="complement">for possession of marijuana</phr>
</phr>
</phr>
</cl>
</cl>
</s>
```

Et, finalement, il faut rappeler cette évidence: la TEI fournit des conventions d'encodage d'un texte mais n'est pas un outil de manipulation de ces textes encodés. La TEI ne "fait" rien. Le linguiste désireux d'exploiter son corpus - ou même simplement de lui donner une forme humainement lisible - devra donc se familiariser avec plusieurs outils complémentaires dont la maîtrise, de nouveau, est loin d'être immédiate: par exemple XPath ou XQuery pour conduire des recherches dans ce corpus, XSLT pour en produire des versions humainement lisibles. L'entreprise est terriblement chronophage.

Conclusions

Dans les conditions usuelles du travail d'un linguiste ordinaire, dans la phase d'enregistrement et de traitement de son corpus, mieux vaut renoncer à une approche comme celle de la TEI, aussi séduisante puisse-t-elle paraître. Ou alors il faudra se limiter à une structuration très simple. Par exemple la transcription et les métadonnées. Une telle approche aurait pleinement sa place dans une perspective qui s'est beaucoup développée depuis une quinzaine d'années, en raison du climat d'urgence généré par le discours sur les "langues en danger", celle d'une "documentation linguistique"¹⁶ qui travaille au recueil de productions langagières variées sans que la préoccupation descriptiviste soit forcément immédiatement présente. Mais cette perspective n'est pas la nôtre ici.

Le linguiste ordinaire devrait-il donc renoncer au précieux secours de l'informatique? Certainement pas. Il y a des alternatives plus légères et mieux adaptées à son travail – constitution d'un corpus annoté et exploitation de celui-ci – qu'un encodage conforme à la TEI. Le texte brut du corpus - sa transcription - doit, bien entendu, se présenter sous la forme d'un fichier informatique. Ensuite on investira utilement du temps dans la maîtrise d'un tableur¹⁷ ou d'un logiciel spécialisé dans la gestion des corpus linguistiques¹⁸ et, surtout, dans celle du langage des expressions régulières qui permet de conduire de nombreuses recherches à partir des propriétés formelles des données textuelles.

1- La préparation, durant mon séminaire de Linguistique théorique et expérimentale de 2012-2013, d'un schéma d'encodage d'un corpus de kabyle, conforme à la TEI et incluant une analyse linguistique.

2-La définition de François Rastier ("un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.", *Enjeux épistémologiques de la linguistique de corpus*, 2002, http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html) est intéressante mais l'adopter c'est

réduire singulièrement l'acception usuelle du terme. L'un des corpus les plus célèbres, et une brillante réussite de la TEI (cf.

<http://www.natcorp.ox.ac.uk/docs/URG/>), le British National Corpus, ne serait pas un corpus au sens de Rastier. Il n'est pas constitué de textes intégraux.

3- Par exemple celui de Bernard Comrie et Norval Smith: *Lingua Descriptive Studies: questionnaire*, *Lingua*, 42, 1, 1977. Cf. <http://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/linguaQ.php>

4- Denise François, *Français parlé, Analyse des unités phoniques et significatives d'un corpus recueilli dans la région parisienne*, Paris, Selaf, 1974, 2 vol., 842 p. Les conventions adoptées pour cette transcription sont exposées aux p. 50-55.

5- Cf. les ressources déposées sur le site du LACITO (laboratoire CNRS-Paris 3 Langues et civilisations à tradition orale)

http://lacito.vjf.cnrs.fr/archivage/tools/list_rsc.php?lg=Ubykh&aff=oubykh

6- On retrouve la distinction entre analyse et présentation explicitée par André Martinet dans *Analyse et présentation, deux temps du travail du linguiste*, in: Jean Dierickx, Yvan Lebrun, *Linguistique contemporaine, Hommage à Eric Buysens*, Université libre de Bruxelles, Bruxelles, 1970, p. 133-140, repris dans: André Martinet, *Studies in FunctionalSyntax*, Wilhelm Fink, München, 1975, p. 134-141 et 267-268.

7- Elan est développé par le Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. Cf. <http://tla.mpi.nl/tools/tla-tools/elan/>

8- http://lacito.vjf.cnrs.fr/archivage/tools/show_text.php?id=crdo-UBY_POISSON_SOUND&id_ref=crdo-UBY_POISSON

9- Cf. <http://www.tei-c.org>

10- Pour un exemple cf. : Alexandre Moatti, Bibliothèque numérique européenne: de l'utopie aux réalités, *Réalités industrielles*, novembre 2012. Il est vrai qu'avec la TEI on se situe en amont de toute application informatique concrète.

11 - Voir la liste de discussion tei-fr@groupe.renater.fr ; en anglais: tei-l@listserv.brown.edu

12- Il s'agit d'Oxygen (<http://www.oxygenxml.com/>). Il reste que la maîtrise de ce logiciel n'est pas évidente si l'on n'est pas familiarisé avec un environnement de programmation. Et sa documentation, en anglais, n'est pas un modèle de pédagogie.

13 - Je laisse de côté les critiques radicales (omniprésence d'une conception hiérarchique des relations entre éléments descriptifs; adjonctions au texte embarquées, incluses, dans le texte lui-même). Elles sont très pertinentes mais conduiraient à écarter toute forme d'encodage du genre de la TEI et à mettre ainsi fin à la discussion. Cf. Desmond Schmidt, *The inadequacy of embedded markup for*

cultural heritage texts, *Literary and Linguistic Computing*, 25, 3, 2010, p. 337-356

14-Sauf si des procédures automatiques sont envisageables. Mais alors à quoi sert l'étiquetage par des balises si l'information peut être retrouvée automatiquement sans elles?

15- Cf. Guidelines, 17.1.1 <http://www.tei-c.org/release/doc/tei-p5-oc/en/html/AI.html>

16 - Nikolaus P. Himmelmann, Documentary and descriptive linguistics, *Linguistics*, 36, 1998, p. 161-195

17- Excel, Calc, Gnumeric etc. L'intérêt de l'utilisation d'un tableur pour une étude linguistique est souvent méconnue ou sous-estimée. Michel Lemaire en a donné des exemples orientés vers une analyse littéraire mais qui peuvent donner des idées. Cf. <http://www.uottawa.ca/academic/arts/astrolabe/articles/art0050/Tableaux1.htm>

18- ELAN, <http://tla.mpi.nl/tools/tla-tools/elan/>; EXMARaLDA, http://www.exmaralda.org/en_downloads.html; CLAN, <http://childes.psy.cmu.edu/klan/> etc.