

To build a chunker

*Implémentations d’un parser de chunks*



BOURDILLAT Jérémy

Université Grenoble Alpes

M1 Sciences du Langage parcours Industries de la Langue

Table des matières

[Introduction : Les *chunks*, une réalité psycholinguistique 2](#_Toc134526296)

[Protocole : Texte de référence et *Gold Standard* 4](#_Toc134526297)

[Protocole et objectifs 4](#_Toc134526298)

[Texte de référence 4](#_Toc134526299)

[Annexes 5](#_Toc134526300)

[Texte de référence 5](#_Toc134526301)

[Segmentation de référence 6](#_Toc134526302)

[Bibliographie 8](#_Toc134526303)

# Introduction : Les *chunks*, une réalité psycholinguistique

« *I begin with an intuition* ». Tels étaient les premiers mots d’Abney, non à sa naissance (encore que) mais dans son article Parsing By Chunks de 1991. L’intuition à laquelle se réfère l’universitaire est une intuition orale, celle d’un schéma prosodique ressenti par le locuteur (Abney, « Parsing By Chunks », 1). Ce schéma prosodique est, de plus, spécifique à l’anglais puisque basé sur son système d’accentuation (un accent lexical par chunk, et seuls les mots lexicaux portent l’accent prosodique en neutral tonicity).

Un chunk est donc un segment ; le plus souvent, il s’agira d’une segmentation de la phrase, c’est-à-dire un découpage des unités dans les limites de cette dernière. Abney déjà les décrivait comme une unité ayant une structure et s’inscrivant dans une structure plus grande : « *I assume that a chunk has syntactic structure which comprises a connected subgraph of the sentence’s parse-tree* » (Abney, p.2). Blache les décrit comme des structures partielles assimilables à des structures syntaxiques mais décrites différemment : il considère un syntagme comme un ensemble de propriétés qui, sous une certaine combinaison, activent une catégorie en particulier plus qu’une autre (Blache, p.231). Il ne s’agit donc pas, d’après lui, de les décrire avec des liens syntaxiques traditionnels.

Or, le découpage en segments est presque toujours effectué sur de la langue écrite ou bien transcrite ; et dans le cas de l’étude du français, dont le système accentuel est différent de celui de l’anglais, doit-on abandonner pour autant cette idée ?

Se pose alors la question de la légitimité d’un tel formalisme dans ce cadre-ci.   
Il s’avère que des études subséquentes du début du millénaire tendent à prouver la validité d’un même phénomène à l’écrit, en relevant que, dans l’élaboration de théories cognitivistes, le découpage en chunks est intéressant d’un point de vue *mémoire* car il permet son optimisation dans le traitement cognitif de la langue (Blache, p.231), sans distinction écrit/oral, via un système de mémoires tampons. D’ailleurs, en sciences cognitives, le *chunk* est une « unité d’information » qui peut être gardée en mémoire à court terme, avec une moyenne définie par George Miller à 7 buffers pour un être humain, appelé « empan mnésique » (L’Homme cognitif, p.343). On peut alors supposer un isomorphisme entre le *chunk* cognitif, et le *chunk* linguistique. Le cognitivisme étant une approche de la pensée comme traitement de l’information, il n’est pas ridicule de stipuler un lien entre informatique, cognition et langage.

Cependant, l’utilité d’un traitement automatique par chunk n’apparaît pas tout de suite clairement : si effectivement l’être humain pourrait fonctionner de cette manière, où réside l’intérêt dans le fait de l’imiter ?

On peut trouver un début de réponse à cette question en relevant le fait que, dans l’hypothèse stipulée par Philippe Blache d’un « traitement incrémental du langage » (*Blache, « Chunks et activation : un modèle de facilitation du traitement linguistique », 230*), on peut mesurer la quantité d’information linguistique apportée par un mot via les différents traits sémantiques, syntactiques ou autres qu’il porte, restreignant ainsi « l’espace de recherche », soit la combinatoire de ce qui peut suivre. Sous cet angle de théorie de l’information, on se rapproche ainsi des travaux de Blache sur les grammaires de propriétés et l’intérêt de fonctionner par contraintes. Tout prend alors sens : en découpant le texte en unités plus petites ayant une charge sémantico-syntaxique (bref, de l’information quantifiable, les *propriétés* (Blache, *Les grammaires de propriétés*, 107.)), on réduit les chemins possibles, et donc l’ambiguïté. De plus, l’apport d’informations et la combinaison de ces dernières peut permettre d’appliquer de nouvelles contraintes : un déterminant peut dénoter un syntagme nominal (ou ce qui s’en rapproche), or s’il est suivi d’une préposition (donc possible syntagme prépositionnel) on peut émettre l’hypothèse d’une relation « complément du nom » entre les deux (entre autres). En suivant la théorie de Blache, obtenir un nom permet de satisfaire des propriétés de constituance, d’obligation et d’unicité vis-à-vis d’un potentiel syntagme nominal. Nous y reviendrons, mais en définissant une hiérarchie claire des possibles (et donc un choix évident, fût-il correct ou non) cela permet une certaine robustesse en écartant la combinatoire, et permet aussi un traitement par point fixe (*fixed-point computation*) (Cooper, *Engineering a Compiler*, 53) souvent utilisé dans les moteurs d’inférence et a fortiori, les compilateurs.

Il est à noter que la théorie de Blache va encore plus loin : ce dernier rapporte le nombre de propriétés présentes au nombre de celles possibles, dites « pertinentes », faisant de son formalisme un hybride de système expert et statistique (Blache p.237).

# Protocole : Texte de référence et *Gold Standard*

## Protocole et objectifs

Nous souhaitons donc réaliser une ébauche de chunker, soit un système qui va ordonner des unités contigües en groupes, avec :

* Pour unités, des **tokens** en français
* Pour règle générale de groupement, une approximation des **syntagmes** de la grammaire traditionnelle

On souhaite aussi construire une structure arborescente, assez proche d’une analyse en constituants, dont la sortie lisible éventuelle sera au format XML.

Il s’agira d’un **moteur d’inférence**, c’est-à-dire un système qui s’appuie sur la logique de première ordre (notions de variables mises en relation par des prédicats) pour prendre des décisions sur la segmentation et la catégorisation des chunks. Notre ordinateur étant une machine logique avant tout, cela en est d’autant plus facile à implémenter.

## Texte de référence

Pour à la fois se donner une direction à prendre et évaluer les performances de nos essais d’implémentation, nous avons choisi un texte de référence (p.5) et défini une segmentation de référence (p.6, seulement au premier niveau), disponibles en annexe.

La segmentation de référence a été réalisée au fur-et-à-mesure du cours, elle ne suit donc pas strictement un paradigme en particulier.

## Evaluation

Les résultats sont évalués à la main, en comparant chaque chunk sorti par le programme avec la référence. Une note est alors attribuée, de 0 à 2 avec 0 : insatisfaisant et 2 : très satisfaisant. Sont évalués le type de chunk et sa constitution (quels tokens ont ou n’ont pas été ajoutés) ; il est possible d’avoir un résultat satisfaisant même si différent quand le chunk résultant porte une charge sémantique ou syntaxique suffisante, à l’appréciation du validateur (une jolie façon de dire que j’ai fait au *feeling*).

# Mise en forme de la base de connaissances

Tout moteur d’inférence possède une base de connaissances, puisque c’est ce qui lui permet de tirer des conclusions. Les raisonnements effectués seront des

# Annexes

### Texte de référence

|  |
| --- |
| Les signaux étaient assez explicites mais Jérémy espérait se tromper. “ J’avais lu sur Doctissimo que cela faisait partie des symptômes mais je n’ai pas voulu m’alarmer. Les gens se moquaient de moi et me disaient mais non, t’en fais pas, t’es pas Français, t’es juste un peu surmené en ce moment.” témoigne-t-il alors qu’il râle dans les embouteillages.  Mais malheureusement, hier, le couperet tombe, après une prise de sang, Jérémy est positif à la nationalité française, une affection longue durée très handicapante remboursée à 100% par la Sécurité sociale. “ Toute votre vie bascule en un instant, vous savez qu’il y a des choses que vous ne pourrez plus jamais apprécier : un voyage au soleil, un film, un mariage. Vous trouverez toujours quelque chose qui ne va pas. Vous savez que quand on vous demandera “ça va “ vous répondrez désormais des phrases comme “ écoute, on fait comme on peut” nous glisse le primo Français qui broie du noir depuis qu’il a vu la météo ce matin.  Il n’existe aucun traitement pour soigner cette terrible maladie alors que près de 65 millions de personnes sont atteintes de nationalité française. “ Pour l’instant la seule chose qui marche pour les soulager, c’est le vin. Mais ce n’est qu’un soin palliatif qui ne traite pas la cause de la maladie ” déclare le docteur Bernard Moutier lui-même atteint par ce terrible mal. |

1 https://www.legorafi.fr/2022/05/25/deprime-et-jamais-content-un-homme-de-37-ans-se-voit-diagnostiquer-francais/

### Segmentation de référence

|  |
| --- |
| Les signaux  étaient  assez explicites  mais  Jérémy  espérait  se tromper  .  “  J’avais lu  sur Doctissimo  que  cela faisait  partie  des symptômes  mais  je n’ai pas voulu  m’alarmer  .  Les gens  se moquaient  de moi  et  me disaient  mais  non  ,  t’en fais pas  ,  t’es pas  Français  ,  t’es juste  un peu  surmené  en ce moment  .  ”  témoigne-t-il  alors qu’  il râle  dans les embouteillages  .  Mais  malheureusement  ,  hier  ,  le couperet  tombe  ,  après une prise  de sang  ,  Jérémy  est positif  à la nationalité française  ,  une affection  longue durée  très handicapante  remboursée  à 100%  par la Sécurité sociale  .  “  Toute votre vie  bascule  en un instant  ,  vous savez  qu’  il y a  des choses  que  vous ne pourrez plus jamais apprécier  :  un voyage  au soleil  ,  un film  ,  un mariage  .  Vous trouverez toujours  quelque chose  qui ne va pas  .  Vous savez  que  quand  on vous demandera  “  ça va  “  vous répondrez  désormais  des phrases comme  “  écoute  ,  on fait  comme  on peut  ”  nous glisse  le primo Français  qui broie  broie  du noir  depuis  qu’  il a vu  la météo  ce matin  .  Il n’existe  aucun traitement  pour soigner  cette terrible maladie  alors que  près de 65 millions  de personnes  sont atteintes  de nationalité française  .  “  Pour l’instant  la seule chose  qui marche  pour les soulager  ,  c’est  le vin  .  Mais  ce n’est  qu’  un soin palliatif  qui ne traite pas  la cause  de la maladie  ”  déclare  le docteur  Bernard Moutier  lui-même atteint  par ce terrible mal  . |

2 texte de référence chunké à la main

# Bibliographie