Meta
Journal des traducteurs
Translators' Journal

META

Construction d'un dictionnaire : morphologie à deux niveaux pour le français à l'aide de contraintes basées sur les structures de traits typées

Fiammetta Namer et Paul Schmidt

Volume 42, numéro 1, mars 1997

Lexicologie et terminologie

URI : https://id.erudit.org/iderudit/003695ar
DOI : https://doi.org/10.7202/003695ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Namer, F. & Schmidt, P. (1997). Construction d'un dictionnaire : morphologie à deux niveaux pour le français à l'aide de contraintes basées sur les structures de traits typées. *Meta*, *42*(1), 72–93. https://doi.org/10.7202/003695ar

Résumé de l'article

Cet article illustre l'application de techniques modernes en linguistique computationnelle et génie linguistique, connues sous le nom de "morphologie à deux niveaux" (MON), "linguistique basée sur les contraintes" et "structures de traits typés" (SST), pour la construction de lexiques en français. Ces techniques s'inspirent de concepts présentés dans la littérature consacrée aux grammaires syntagmatiques guidées par les têtes (HPSG) et à la MON pour le traitement de la morphologie flexionnelle du français. Le formalisme dans lequel le lexique a été réalisé est ALEP (advanced language engineering platform), conçu pour associer expressivité et efficacité. Les deux modules intervenant dans la construction du lexique sont le composant à deux niveaux et le composant morphosyntaxique. Leur description ainsi que leurs rôles respectifs sont présentés et illustrés en détail.

CONSTRUCTION D'UN DICTIONNAIRE : MORPHOLOGIE À DEUX NIVEAUX POUR LE FRANÇAIS À L'AIDE DE CONTRAINTES BASÉES SUR LES STRUCTURES DE TRAITS TYPÉES*

FIAMMETTA NAMER ET PAUL SCHMIDT
TALANA — Université Nancy II, Nancy, France et IAI, Saarbrücken, Allemagne

Résumé

Cet article illustre l'application de techniques modernes en linguistique computationnelle et génie linguistique, connues sous le nom de «morphologie à deux niveaux» (MDN), «linguistique basée sur les contraintes» et «structures de traits typés» (SST), pour la construction de lexiques en français. Ces techniques s'inspirent de concepts présentés dans la littérature consacrée aux grammaires syntagmatiques guidées par les têtes (HPSG) et à la MDN pour le traitement de la morphologie flexionnelle du français. Le formalisme dans lequel le lexique a été réalisé est ALEP (advanced language engineering platform), conçu pour associer expressivité et efficacité. Les deux modules intervenant dans la construction du lexique sont le composant à deux niveaux et le composant morphosyntaxique. Leur description ainsi que leurs rôles respectifs sont présentés et illustrés en détail.

Abstract

In this article, we demonstrate the application of modern techniques used in computational linguistics and linguistic engineering for compiling lexicons in French. Known as two-level morphology, constraint-based linguistics *and* typed feature structures, *these techniques are derived from concepts dealt with in texts on HPSG and flexional two-level morphology. ALEP (Advanced language engineering platform), a system designed to provide maximum expressiveness and efficiency, was used. Its two components, one two-level, the other morpho-syntactical, are fully described.*

1. INTRODUCTION

Un aspect important, dans la conception d'un dictionnaire électronique destiné au traitement automatique du langage, consiste à éviter la redondance d'informations. Construire un dictionnaire de formes non fléchies (en fait, un dictionnaire de morphèmes) signifie disposer d'un outil morphologique efficace et réversible qui génère et manipule des structures de mots généralisées.

Nous proposons de présenter la définition d'un tel composant, c'est-à-dire les mécanismes utilisés pour la manipulation d'informations morphologiques pour le français (limitées pour l'instant aux informations flexionnelles). Ces mécanismes sont fondés sur des techniques linguistiques développées récemment, comme la MDN (morphologie à deux niveaux), et s'inspirent des travaux de Koskiennemi (1983), Trost (1990) et Ruessink (1990), ou comme le concept de STT (structures de traits typées), qui est à la base du formalisme grammatical HPSG (grammaire syntagmatique guidée par les têtes, cf. Pollard et Sag (1987), Pollard et Sag (1994) et Krieger (1994)). La MDN prend en compte les variations orthographiques entre chaîne de surface et chaîne lexicale, effectuant ainsi une

segmentation efficace. La notion de STT est utilisée pour la description et l'interprétation des combinaisons morphosyntaxiques[1].

Le cadre formel qui est à la base du travail présenté est ALEP (*Advanced Language Engineering Platform*), un système qui a été développé pour le génie linguistique à grande échelle (cf. Alshawi *et al.*). Ce système a été conçu pour être efficace, ce qui signifie qu'il n'inclut que des mécanismes performants du point de vue de leur implémentation informatique, donc essentiellement des composants qui peuvent être directement compilés en Prolog.

Ce formalisme fournit les mécanismes requis pour le module morphologique, que nous présentons dans cet article :

- Un puissant outil à deux niveaux aborde les mêmes problèmes que ses prédécesseurs, et permet d'effectuer toutes les opérations nécessaires à la manipulation des variations orthographiques. Cet outil a été appliqué au traitement d'un grand nombre de phénomènes, dans plusieurs langues. Ici, il est appliqué aux problèmes orthographiques de la flexion du français ; citons le doublement de la consonne dans certaines formes adjectivales au féminin comme `bonne`. Le but est de ne conserver dans le lexique que la base `bon` et la terminaison `e`. Le découpage, ainsi que le traitement de la variation orthographique représentée ici par le doublement de la consonne, constituent les problèmes typiques de la transcription graphème-morphèmes : le travail du formalisme à deux niveaux consiste à permettre (et contraindre) la segmentation de la chaîne de surfaces `bonne` en ses deux items lexicaux `bon+e`.
- Un composant morphosyntaxique permet de définir la structure des mots. Une fois que `bonne` a été segmenté en `bon+e+`, il reste à établir que cette combinaison de morphèmes est licite, qu'elle correspond à une structure adjectivale, au féminin-singulier. En relation avec ce composant, les points suivants sont abordés dans cet article :

  1. adoption de l'approche structurelle de la morphosyntaxe, *i.e.* à l'aide de règles de la forme **mot ⇒ base suffixe**, où chaque item est un objet structuré complexe représenté par une STT. Nous n'adoptons ni l'approche basée sur les règles lexicales (Pollard et Sag 1987 : 191-218), ni l'approche dite mot-et-paradigme (Krieger 1994) ;
  2. description des notions de «tête» et de «sous-catégorisation morphologique» ;
  3. structuration de l'information lexicale morphologique dans un cadre basé sur les STT inspirés de HPSG, qui définit une interface avec la syntaxe et la sémantique.

Ces concepts seront appliqués pour donner une description complète et élégante de la flexion du français.

2. MORPHOLOGIE À DEUX NIVEAUX (MDN)

2.1. LE FORMALISME À DEUX NIVEAUX EN ALEP

2.1.1. Introduction

Les concepteurs du formalisme à deux niveaux d'ALEP ont tenu compte des principaux problèmes que les précédents formalismes à deux niveaux (Koskiennemi 1983 ; Black 1987, etc.), ont abordés sans toujours réussir à les résoudre.

- Les règles d'appariement surface / lexique doivent permettre la spécification des contextes gauche et droit pour les deux chaînes mises en correspondance. Elles doivent également permettre la mise en relation de séquences ayant des longueurs arbitraires. Ainsi, la forme générale d'une règle est la suivante :

```
<ContexteSurfGauche> <SG> <ContexteSurfDroit> Operateur
<ContexteLexGauche> <SD> <ContexteLexDroit>
```

Chaque variable, représentée entre «< >», instancie une liste de caractères.

Ainsi, si l'on veut traiter la formation irrégulière du féminin par doublement de la consonne finale de la base, comme celle présentée à la section 1 :

(1) bonne ⇛ bon +e
base e_{fem}

il faut (a) imposer l'identité entre le contexte gauche de surface (`ContexteSurfGauche`) et la séquence gauche (`SG`), *i.e.* la consonne finale «n» de la base «n», et (b) contraindre le contexte droit de surface (`ContexteSurfDroit`) à la terminaison du féminin. La séquence droite (`SD`) est constituée par le symbole de fin de morphème «+». La valeur des contextes n'est pas précisée dans la partie droite :

(2) [n] [n] [e] ⟶ [] [+] []

L'analyse d'une chaîne donnée réussit s'il y a au moins une règle de mise en correspondance pour chaque séquence constituant la chaîne de surface et/ou lexicale.

■ Deux types d'`Opérateurs` sont requis : «⇛» signifie que l'appariement est optionnel, *i.e.* la même `SG`, avec les mêmes contextes, peut être mise en correspondance avec une autre `SD`. « ⟸⟹ » indique, lui, un appariement obligatoire, *i.e.* il existe une et une seule mise en correspondance entre surface et lexique. Bien entendu, ces opérateurs agissent non seulement en analyse (surface vers lexique) mais aussi en génération (lexique vers surface).

Ainsi, pour représenter un phénomène tel que (1), l'opérateur choisi est ⟸⟹, qui exclut l'application de toute autre règle.

■ L'application des règles doit être par ailleurs contrainte au moyen d'informations linguistiques codées dans les entrées lexicales des morphèmes pertinents, pour caractériser ceux dont le comportement est compatible avec la règle. Ainsi, pour (2), la STT :

(3) $_{astem}\left[\text{DOUBLE oui}\right]$

indique que la séquence lexicale affectée par la règle doit appartenir à une base adjectivale (*astem*) qui est caractérisée par le doublement de sa consonne finale au féminin.

■ Le formalisme autorise la définition d'ensemble des caractères, et la manipulation de variables sur ces ensembles. Ainsi, il est possible d'exprimer dans (4), au moyen de la variable Y, que la séquence «nne», proposée dans (2) correspond à la fin du mot (symbole «=»[2]) ou peut être suivie du marqueur «s» du pluriel. D'autre part, la paramétrisation de `SG` et de `ContexteSurfGauche`, au moyen de la variable C, permet d'étendre le champ d'application de la règle à d'autres cas de doublement de consonnes («bas/basse», «nul/nulle»,...) :

(4) [C] [C] [e,Y] ⟸⟹ [] [+] [],
$_{astem}\left[\text{DOUBLE oui}\right]$,
[C in { n, s, l}, Y in {s,=}]

(4) est presque la représentation d'une règle ALEP de MDN : il ne reste plus qu'à ajouter un identificateur («double_cons») et à interpréter l'ensemble des items mentionnés jusqu'ici au moyen du terme Prolog à trois places «tlm_rule». Le résultat, (5) est alors une description ALEP à deux niveaux :

(5) tlm_rule(
[C] [C] [e,Y] ⟸⟹ [] [+] [],
astem [DOUBLE oui],
[C in { n, s, l}, Y in { s,=}]
).

C'est à l'aide de ce dispositif que nous présentons ci-dessous les principaux aspects des variations orthographiques flexionnelles des verbes, noms et adjectifs du français.

2.1.2. Techniques de base

Le formalisme MDN proposé est un mécanisme très puissant, dont l'utilisation abusive peut entraîner une baisse sensible d'efficacité. Il faut donc déterminer les limites optimales d'utilisation du module MDN, *i.e.* décider quelle est la division des tâches souhaitée entre MDN et morphosyntaxe. Il y a deux cas extrêmes.

1. Tout est fait dans le lexique et dans le composant morphosyntaxique. Cela implique que le module de MDN est simplement utilisé en tant que mécanisme d'appariement qui met en correspondance une chaîne de surface avec une chaîne lexicale identique. Le seul usage que l'on puisse faire d'une telle application du formalisme à deux niveaux est la segmentation entièrement compositionnelle des mots en morphèmes. Trois règles sont alors suffisantes : la première apparie le symbole de fin de mot au symbole de fin de morphème, qui, en ALEP, sont respectivement les caractères «=» et «+». Les deux autres règles sont plus intéressantes :

La règle `segment` tente d'insérer un symbole de fin de morphème entre chaque caractère. La validité de chaque séquence résultante est vérifiée dans le lexique. La règle `défaut` rend compte des appariements par défaut : un caractère de surface est transcrit en caractère identique dans le lexique :

(6)
```
tlm_rule(                          tlm_rule(
segment,                           defaut,
[] [] [] ⇒ [] [+] [],              [] [X] [] ⇒ [] [X] [],
).                                 ).
```

2. L'autre extrême consiste à tout faire dans le composant à deux niveaux. Cela revient, par exemple, à considérer des formes verbales totalement irrégulières comme autant de cas de variations orthographiques :

(7)
```
tlm_rule(
défaut,
[] [p,r,i,s] [=] ⟸⟹ [] [p,r,e,n,d,+,i,s] [+],
).
```

Aucune des deux solutions n'est souhaitable : la première implique un lexique dans lequel toutes les variations orthographiques sont représentées, et un module de MDN totalement non déterministe. La seconde, bien qu'entraînant la réduction du lexique, se caractérise par un module de MDN difficile à manier, impossible à généraliser, à maintenir et à contrôler.

En fait, le problème de la division des tâches entre MDN et morphosyntaxe consiste en deux questions :

1. Quand est-ce qu'une forme doit être décomposée et non pas intégrée dans le lexique ?
2. Quand une forme est considérée comme décomposable, quelle technique utiliser en MDN ?

Suivant Ritchie *et al.* (1992), qui proposent une réponse générale à la première question, un critère de «décompositionnalité» pour un mot est fonction de ce qui est «raisonnablement» facile à faire, *i.e.* des potentialités de ce mot à formuler des règles reflétant une certaine généralité.

Étudions maintenant plus en détail les réponses possibles à la seconde question: Quelles techniques de MDN utiliser pour la segmentation et le traitement des variations orthographiques de mots dont la forme est «raisonnablement» compositionnelle et générale?

Les cas typiques d'application de MDN s'observent quand la compositionnalité est légèrement perturbée. Citons ici deux classes majeures de phénomènes communs à plusieurs langues européennes. Des données spécifiques au français sont présentées dans la section 2.2.:

- Suppression/insertion de caractères:
  1. doublement de consonne (cf. (3));
  2. cas du «e» muet.

  Les règles de mise en correspondance doivent générer le caractère supplémentaire, soit dans le lexique, soit en surface comme c'est le cas dans l'exemple (3).
- Changement de caractère:
  1. variation de consonne (`pouv+oir` vs `pour+ra`);
  2. décalage de voyelle dans les langues germaniques;
  3. variations d'ordre phonologique, comme l'alternance y/i en français dans (`essu[y]-er` vs `essu[i]-ent`).

Il n'y a aucun doute que de tels phénomènes appartiennent à la MDN. La technique appliquée à la première classe consiste à supprimer un caractère dans certaines conditions, comme le montre (3).

La seconde classe de phénomènes sous-entend un changement de caractère. Le problème est alors de contrôler l'application des règles qui modifient ce caractère. Trost (1990) propose l'utilisation de diacritiques, que nous présentons brièvement ici, au moyen de l'exemple de l'alternance y/i. D'autres exemples, proposés dans la section 2.2.2., illustrent cette approche en détail.

Le morphème représentant la base des verbes qui subissent l'alternance y/i est représenté dans le lexique avec un caractère, non déclaré dans le lexique, qui neutralise les caractères qui alternent. Ainsi, «Y» représente arbitrairement l'alternance y/i dans la conjugaison du verbe `essuyer`: [essuY+]. Deux règles sont nécessaires au traitement de l'alternance. Ainsi, pour `essuient/essuyer`:

(i) La règle (8) apparie «y» à «Y» lorsque la séquence correspondante ne donne pas lieu à une variation y/i: quand la chaîne de surface contient le caractère «y», le morphème, dans le lexique, est caractérisé par le trait [Y-I-ALTERNANCE non].

(8)
```
tlm_rule(
    y-i-alternance_non,
    [] [y] [] <==> [] ['Y'] [],
    morph [Y-I-ALTERNANCE non] ).
```

(ii) L'autre règle met en correspondance «i» avec «Y», quand le morphème est affecté par la variation y/i, *i.e.* lorsque la SG est un «i»:

(9) tlm_rule(
y-i-alternance_oui,
[] [i] [] ⟺ [] ['Y'] [],
morph [Y-I-ALTERNANCE oui]).

La notation diacritique par «Y» rend impossible le traitement de la base par la règle d'appariement par défaut (7) et en garantit une application contrôlée par les deux règles ci-dessus.

De manière à désactiver ces deux règles pour tout morphème non concerné par l'alternance y/i (p. ex. `grasseyer`), le trait Y-I-ALTERNANCE peut prendre la valeur «aucun» : cet attribut est ainsi instancié à cette valeur chaque fois que le phénomène d'alternance y-i est non pertinent.

À l'inverse, les morphèmes comme «essuY» reçoivent le trait [Y-I-ALTERNANCE ¬ aucun], dont la valeur est interprétable comme «oui» ou «non», ce qui permet l'activation des règles (8) et (9).

Du point de vue de l'analyse, une fois que la formulation négative [Y-I-ALTERNANCES¬ aucun] a été spécifiée à «oui» ou «non» par la règle de MDN qui rend compte de la variation y-i, c'est au tour de la morphosyntaxe d'interpréter correctement l'occurrence ou la non-occurrence de «i» ou de «y». Ainsi, dans le cas de *essuient/essuyer*, le suffixe «ent» doit être marqué comme se combinant avec une base [Y-I-ALTERNANCE ¬ non], alors que le suffixe «er» doit se combiner avec une base [Y-I-ALTERNANCE ¬ oui].

Outre une illustration détaillée des deux classes de phénomènes orthographiques en français, la section ci-dessous présente une troisième classe de phénomènes qui n'ont rien à voir avec la compositionnalité, mais dont le traitement en MDN permet d'augmenter l'efficacité du module morphologique en diminuant de façon drastique la taille du lexique des suffixes flexionnels du français.

2.2. LE COMPOSANT À DEUX NIVEAUX POUR LE FRANÇAIS

Le français est une langue dont les paradigmes flexionnels sont nombreux et complexes (cf. Bescherelle, 1990). Les propriétés du formalisme d'ALEP pour la MDN nous ont permis de mettre en place de façon efficace un système de correspondances entre les formes fléchies du français et un nombre minimal de morphèmes. Avant de présenter les mécanismes mis en œuvre pour la construction du dictionnaire des morphèmes, nous examinons brièvement les données flexionnelles du français, ainsi que les variations orthographiques correspondantes.

2.2.1. Données

Dans un mot fléchi, on distingue la base, qui lui confère ses propriétés intrinsèques, relationnelles et sémantiques, et la terminaison, constituée de suffixes pouvant apporter chacun une ou plusieurs informations flexionnelles — éventuellement contradictoires — à l'ensemble. Certaines familles de base, aussi bien que de suffixes, peuvent être regroupées en classes d'allomorphes, suivant le type de variations orthographiques ou phonologiques que ces séquences subissent durant leur paradigme flexionnel. L'étude de ces variations met en évidence deux types distincts de phénomènes : ceux qui affectent les bases et ceux qui affectent les suffixes.

Les bases :

Certaines bases de mots sont soit totalement stables (*aim-*, *petit-*, *enfant-*), soit totalement irrégulières (*all(er)/ir-(ai)/v-(a)*, *œil/yeux*, *vieux/vieille*). Dans les deux cas, les règles de mise en correspondance graphème-morphème sont inutiles : autant d'entrées

lexicales que de bases sont intégrées dans le lexique. Le troisième cas de figure, que nous examinons plus en détail, concerne des phénomènes orthographiques localisés et réguliers. La nature de ces phénomènes est indépendante de la catégorie grammaticale de la base : certaines variations phonologiques ou purement orthographiques concernent uniquement les verbes, ou les noms, ou les adjectifs, mais en général au moins deux catégories sur trois sont affectées, comme nous allons le voir.

Phonologie Certaines variations orthographiques relèvent directement de la phonologie :

- conservation de la prononciation d'un phonème, comme avec les variations *g / ge* dans *man[g]-e / man[ge]-ons*, ou *c / ç* dans *lan[c]-e / lan[ç]-ons*, ou encore *e / è* dans *compl[e]t / compl[è]t-e* ;
- variation d'antériorité de la voyelle finale, produisant le schwa, comme dans *act[eu]r / act[]r-ic-e*, *enchant[eu]r*, *enchant[e]r-ess-e*, *ach[è]t-e / ach[e]t-ons*, ou variation d'aperture (*c[é]d-ons / c[è]d-ent*) ;
- alternance (y / i) en relation avec la présence du schwa dans la terminaison : obligatoire, comme pour *essu[y]-er / essu[i]-ent / *essu[y]-ent*[3], optionnelle, comme pour *bala[y]-er / bala[i]-ent / *bala[y]-ent*, ou interdite, comme pour *grasse[y]-er / grasse-[y]-ent / *grasse[i]-ent* ;
- alternance (y / i), cette fois en fonction de la première lettre de la terminaison : *vo[y]-ons / vo[i]-s*, *cro[y]-ant / cro[i]-r-ons* ;
- Doublement de la consonne finale de la base, lorsque le caractère suivant est le schwa : *je[t]-er / je[tt]-e*, *appe[l]-ons / appe[ll]-ent*, *ba[s] / ba[ss]-e*, *nu[l] / nu[ll]-e*, *cocho[n] / cocho[nn]-e*.

Les variations ci-dessus se rangent suivant les deux classes présentées dans la section 2.1.2. : la suppression / insertion d'un caractère (g / g[e], eu / _, etc.), ou l'alternance entre deux caractères (e / é / è, y / i, c / ç, etc.).

La plupart des autres variations orthographiques présentées ci-dessous ont également un lien avec la phonologie, bien que ce lien soit moins évident. Elles se distribuent selon les deux mêmes classes de description :

Insertion / suppression Ce phénomène est très général, comme en témoignent la suppression du «i» final lors de la mise au pluriel de certains noms : *cora[i]l / corau-x*, l'insertion d'une consonne au féminin pour certains adjectifs : *rigolo / rigolo[t]-e*, la suppression de la consonne finale de nombreux verbes au singulier du présent indicatif : *dor[m]-ir / dor[]-t*, ou encore l'insertion d'un caractère au futur : *val-oir / vau[d]-ra*, *cueill-ir / cueill[e]-r-a*.

Alternance Certaines paires de caractères alternent de façon extrêmement régulière, quelle que soit la partie du discours considérée : c'est le cas de u / l dans *anima[l] / anima[u]-x*, *corai[l] / cora[u]-x*, *fina[l] / fina[u]-x*, *abso[l]v-ais / abso[u]d-ra*, *va[l]-oir / va[u]-t*. D'autres alternances concernent par exemple les paires v / d (*absou[d]-re / absol[v]-e*), ou f / v (*bre[f] / brè[v]-e*).

Les suffixes :

Alors que les variations orthographiques des bases n'ont aucune incidence sur les propriétés linguistiques (inhérentes, relationnelles, sémantiques) de celles-ci, les phénomènes qui affectent dans l'ensemble les (très nombreux) suffixes flexionnels du français les identifient selon leur modèle flexionnel, et peuvent être divisés en deux classes.

1. Les suffixes vides du point de vue des informations de temps, aspect, ou accord, et qui ne sont que des marqueurs d'appartenance à une classe flexionnelle. Ainsi, *iss-* permet de distinguer les verbes dits du deuxième groupe de conjugaison de ceux du troisième

groupe, mais ce suffixe n'intervient pas dans la conjugaison proprement dite. De la même façon, le morphème *ess-*, inséré entre la base de certains adjectifs et le suffixe du féminin-singulier *e* (*enchanter-ess-e*) n'apporte à l'ensemble aucune information flexionnelle supplémentaire.

2. Les suffixes informatifs du point de vue flexionnel. Ceux-ci sont beaucoup plus nombreux que les premiers, et peuvent être regroupés en classes d'allomorphes, dont les membres sont représentatifs d'un modèle flexionnel. Ainsi, la marque du futur est assurée par {*er-*, *ir-*, *r-*}, le participe passé variable en nombre est représenté par l'un des caractères {*é*, *i*, *u*, *t*}, alors que les suffixes {*s*, *is*} marquent aussi le participe passé, mais sont invariables en nombre. Citons encore le cas des terminaisons du passé simple, dont la voyelle thématique varie en fonction du modèle de conjugaison : {*a*, *i*, *u*}.

En conclusion, le rôle du système de conversion graphème-morphème doit être triple (outre la segmentation assurée par les trois règles de base, cf. section 2.1.2.) :

- mettre en correspondance les formes allomorphiques de surface des bases avec une chaîne lexicale unique. La mise en correspondance doit être réversible, pour que le lexique puisse s'utiliser aussi bien en analyse qu'en génération ;
- supprimer les suffixes «inutiles» du point de vue morphosyntaxique, tout en gardant une trace de leur présence dans la forme de surface (toujours pour assurer la réversibilité du processus) ;
- «neutraliser» en un seul suffixe les formes allomorphiques porteuses de la même information flexionnelle. Encore une fois, la mise en correspondance doit permettre de générer la forme de surface adéquate, en fonction du modèle de conjugaison auquel le morphème appartient.

Dans ce qui suit, nous avons étudié trois approches possibles pour réaliser ces mises en correspondance dans le formalisme de MDN d'ALEP.

2.2.2. Application des principales stratégies

Les tâches que doit effectuer la grammaire à deux niveaux, résumées ci-dessus, sont des mises en correspondance entre des chaînes qui ont soit la même longueur (alternance entre deux caractères), soit des longueurs différentes (suppression / insertion de caractères). Les exemples ci-dessous illustrent de façon détaillée les techniques présentées en 2.1.2. pour le traitement de l'insertion / suppression — approche dite «alphabétique» (a) — et de l'alternance — approche diacritique (b). Un troisième exemple illustre un cas de paramétrisation des règles de MDN (c).

(a) Approche alphabétique :

Pour illustrer cette approche, prenons l'exemple de la formation irrégulière du féminin pour certains noms et adjectifs, illustrée par les exemples (a)-(d), auxquels s'oppose l'exemple (e) :

(a) *acteur / actrice*,
(b) *maître / maîtresse*,
(c) *enchanteur / enchanteresse*,
(d) *docteur / doctoresse*,
(e) *supérieur / supérieure*.

Le traitement de ces irrégularités peut être distribué sur deux niveaux :

- la base : la mise au féminin entraîne la modification de [eu] en caractère vide (ex. a), ou en [e] (ex. c) ou en [o] (ex. d) ;

■ les suffixes : les suffixes «vides» *ic-* (p. ex. a), *ss-* (p. ex. b) et *ess-* (p. ex. c et d) sont enchâssés entre la base et le suffixe *e*.

En ce qui concerne le problème apparaissant au niveau de la base, les variations phonologiques du son /œ/ sont arbitrairement appariées à la même représentation lexicale, *i.e.* le caractère vide. Ainsi, le lemme identifiant par exemple la base de «acteur/actrice» est `actr`. Un trait est nécessaire pour contraindre l'application de la règle qui supprime les séquences de surface [eu]/[e]/[o] aux exemples (a), (c) et (d). La valeur de ce trait doit être telle que la règle ne s'applique pas à l'exemple (e), puisque la formation du féminin «supérieur-e» est totalement régulière. Par ailleurs, ce trait, appelé EU_SUP(pression), doit faire apparaître la valeur de surface de la séquence variable. Les entrées lexicales résultantes sont les suivantes :

(10) *morph* [LEMME actr, EU_SUP nil] *morph* [LEMME enchantr, EU_SUP e] *morph* [LEMME doctr, EU_SUP o] *morph* [LEMME supérieur, EU_SUP aucun]

Les règles ci-dessous sont activées uniquement pour l'appariement de chaînes identifiant des morphèmes dont l'entrée est caractérisée par le trait [EU_SUP ¬ aucun]. Alors que la règle EU_SUP1 apparie la séquence [eu] à la séquence vide, la règle EU_SUP2 met en correspondance avec le caractère nul la voyelle apparaissant dans la chaîne de surface au féminin. Le caractère de surface est paramétrisé, et sa valeur est contrôlée à la fois dans la liste des conditions et dans le trait EU_SUP :

(11)
thm_rule(
eu_sup1,
[] [e,u] [r,X] ⟺ [] [] [r,+],
morph [EU_SUP ¬ aucun],
[X in {s, = }]).

tlm_rule(
eu_sup2,
[] [O] [r,e,s,s,e] ⟺ [] [] [r,+],
morph [EU_SUP O],
[O in { o, e }]).

L'applicabilité de ces règles s'étend à l'ensemble des morphèmes, ce qui est assuré par le type *morph* introduisant les structures de traits qui contraignent les morphèmes manipulés.

Tournons-nous maintenant vers l'autre problème soulevé par ces exemples, *i.e.* l'insertion de suffixes vides du point de vue de la morphosyntaxe. Comme nous l'avons annoncé, les séquences de surface *ic-* et *(e)ss-* sont supprimées dans la chaîne lexicale, et donc appariées au séparateur abstrait de morphèmes «+». Pour préserver la réversibilité de la règle, un nouveau trait est instancié. Ce trait, appelé SUFF_FEM, spécifie la classe à laquelle appartiennent la base et le suffixe effacé. Le tableau ci-dessous résume les valeurs possibles de ce trait dans les exemples (a)-(e) :

(12)

	LEMME	SUFF_FEM
(a)	`actr`	ic
(b)	`maître`	ss
(c)	`enchantr`	ess
(d)	`doctr`	ess
(e)	`supérieur`	aucun

Les règles de suppression des suffixes sont les suivantes :

(13)
```
tlm_rule(                    tlm_rule(                    tlm_rule
suff_sup1,                   suff_sup2,                   suff_sup3,
[r] [e,s,s] [e,X] <=> [] [+] [],  [e] [s,s] [e,X] <=> [] [+] [],  [r] [i,c] [e,X] <=> [] [+] [],
morph [SUFF_FEM ess],        morph [SUFF_FEM ss],         morph [SUFF_FEM ic],
[ X in {s = } ] ).           [X in { s,= } ] ).           [X in { s,= } ] ).
```

Pour conclure, le tableau ci-dessous récapitule les règles qui sont activées pour mettre en correspondance surface / lexique les adjectifs et noms en -eur :

(14)

SURFACE	REGLE(S)	SEGMENTATION
docteur	eu_sup1	doctr+
doctoresse	eu_sup2, suff_sup1	doctr+e+
enchanteur	eu_sup1	enchantr+
enchanteresse	eu_sup2, suff_sup1	enchantr+e+
maître	defaut	maître+
maîtresse	suff_sup2	maître+e+
acteur	eu_sup1	actr+
actrice	suff_sup3	actr+e+
supérieur	defaut	supérieur+
supérieure	defaut	supérieur+e+

(b) Approche diacritique :

Cette approche est illustrée par un exemple d'alternance de caractères qui affecte les bases verbales, nominales et adjectivales, à savoir l'alternance u/l présentée à la section 2.2.1. Il s'agit d'opposer les paradigmes flexionnels de *animal* et *bal*, *corail* et *attirail*, *final* et *fatal*, *absoudre* et *coudre*. Ainsi, dans l'entrée lexicale des premiers, le caractère de surface variable est neutralisé dans le lexique au moyen du symbole «W». De plus, la différence de comportement de ces deux ensembles de base (l'un affecté par la variation u/l, l'autre régulier de ce point de vue) est marquée par le trait morphologique U-L, dont la valeur est *aucun* pour les bases régulières, alors que cette valeur est niée pour les autres.

Dans ces conditions, les règles `u_to_l_1` et `u_to_l_2` effectuent l'appariement surface / lexique des chaînes caractérisées par le trait [U-L ¬ aucun].

Dans `u_to_l_1`, la mise en correspondance «*l* ↔ 'W'» a lieu si le contexte droit de surface correspond soit à un singulier, soit à la consonne «v» — dans le cas du modèle verbal `absolv-`. De même, `u_to_l_2` apparie «u» à «W» si le contexte droit de surface est approprié :

(15)
```
tlm_rule(                    tlm_rule(
u_to_l_1,                    u_to_l_2,
[] [l] [X] <=> [] ['W'] [],  [] [u] [X] <=> [] ['W'] [],
morph [U-L oui],             morph [U-L non],
[X in {=, v}]).              [X in {x,s,t,d }]).
```

(c) Règles paramétrées :

Les propriétés flexionnelles du français sont telles que les approches alphabétiques et diacritiques sont utilisées le plus fréquemment dans des règles paramétrées, dans lesquelles chaînes et contextes variables sont instanciés par des séquences rendues appropriées au moyen de traits judicieusement choisis. Ces séquences appartiennent à des bases ou suffixes correspondant à des modèles flexionnels différents, mais dont un aspect du paradigme est comparable. Ainsi, on observe qu'une quinzaine de modèles appartenant au troisième groupe de conjugaison est constituée de verbes dont la base perd la consonne finale au singulier de l'indicatif présent (*ser[v]-ir / ser[]-t*; *dor[m]-ir / dor[]-t*; *join[d]-re / join[]-t*,...). On se sert de l'attribut SUPP_PRES pour coder la valeur de cette consonne :

(16) $\underset{vstem}{\begin{bmatrix} \text{LEMME} & \text{serv} \\ \text{SUPP\_PRES} & \text{v} \end{bmatrix}}$ $\underset{vstem}{\begin{bmatrix} \text{LEMME} & \text{dorm} \\ \text{SUPP\_PRES} & \text{m} \end{bmatrix}}$ $\underset{vstem}{\begin{bmatrix} \text{LEMME} & \text{joind} \\ \text{SUPP\_PRES} & \text{d} \end{bmatrix}}$

cette valeur étant représentée sous forme du paramètre «Y» dans la règle (17) d'insertion / suppression. Le contexte droit de surface est la terminaison de l'une des trois personnes du singulier de l'indicatif présent, ce qui est confirmé par le contexte lexical droit, qui indique que les chaînes appariées sont en fin de morphème. Enfin, la chaîne lexicale est la variable Y, qui s'identifie avec la valeur de l'attribut SUPP_PRES. Enfin, le type *vstem* contraint ultérieurement la règle à ne s'appliquer que sur des bases verbales :

(17)
```
tlm_rule(
suppression_present,
[] [] [X]  <==>  [] [Y] [+],
vstem [SUPP_PRES Y],
[X in {s, x, t }]).
```

2.3. RÉSULTATS

Par l'adoption des stratégies décrites et illustrées tout au long de la section 2., l'application du module de MDN entraîne une réduction de la taille du lexique, non seulement pour la représentation des bases, mais également pour celle des suffixes, dont le nombre passe de 34 items distincts (éventuellement ambigus) recensés dans Bescherelle (1990) à un ensemble de 18 morphèmes, dont les formes sont parfois abstraites, afin de supprimer les ambiguïtés.

Le tableau (18) ci-dessous donne l'ensemble des suffixes flexionnels présents dans le lexique.

(18)

a	ai	ant	e	ent	Er	ez	É	i
ons	r	s	S	t	U	Umes	Utes	Urent

Outre la suppression de certaines séquences (*iss-*), le tableau résume l'ensemble des «neutralisations d'allomorphes» effectuées par le module de correspondance graphèmes-morphèmes.

- Une forme unique («Er») factorise les marques de l'infinitif *-er*, *-ir*, *-oir*, *-re*, et se distingue de la chaîne lexicale unique («r») correspondant aux marques du futur / conditionnel, dont certaines se confondent avec l'infinitif : *er-*, *ir-*, *r-*.
- Les différents emplois dc la terminaison «s» (marque de personne vs marque de nombre) sont représentés par les deux entrées : s et S.

■ De la même façon, le suffixe «i», ambigu en surface puisqu'il marque le présent/passé simple indicatif (*fin-i-t*), l'imparfait indicatif (*mang-i-ons*), le conditionnel présent (*mang-er-i-ons*), ou le participe passé (*part-i*) est converti, en fonction des informations qu'il véhicule, en l'un des trois morphèmes distincts : `i`, `U`, `É`.

■ Les morphèmes `Umes`, `Utes` et `Urent` factorisent les variations des terminaisons du pluriel du passé simple dues au groupe de conjugaison.

3. MORPHOSYNTAXE

3.1. INTRODUCTION

Une fois la chaîne de surface convertie en une combinaison base + suffixe(s), le module morphosyntaxique exploite le lexique des morphèmes afin de tester la validité de cette segmentation, et de calculer la valeur des traits morphosyntaxiques, syntaxiques et sémantiques qui en résultent[4]. En d'autres termes, il s'agit de déterminer quelle information est codée dans le lexique, et dans quel type d'entrée lexicale, de façon à éviter toute redondance dans le lexique, et à optimiser l'efficacité du module. La réponse à cette question dépend des données monolingues (cf. section 3.3.) et de l'approche choisie pour leur traitement (cf. section 3.2.). Des exemples d'analyse morphosyntaxique illustrent pour finir les techniques et stratégies employées (cf. section 3.4.).

3.2. STRUCTURE DES MOTS : DIFFÉRENTES APPROCHES

Nous allons nous référer à l'exemple (`essui-ent`) de la section 2.1.2., pour introduire les questions pertinentes dans la construction des descripteurs de la structure des mots. Une fois que `essui-ent` est décomposé, par des règles à deux niveaux, en deux morphèmes, l'un identifié comme base verbale (`essuY`), l'autre (`ent`), comme un suffixe codant la 3^{e} personne du pluriel de l'indicatif présent, ces deux morphèmes doivent être combinés dans une description structurelle qui détermine que le mot `essuient` est un verbe à la 3^{e} personne du pluriel de l'indicatif présent. Clairement, cet exemple montre que le simple traitement hors contexte :

(19) **verbe → base suffixe**

est insuffisant. Il y a en effet deux types de problèmes à résoudre, dont les solutions sont discutées ci-dessous :

■ ceux qui concernent les mécanismes utilisés pour décrire la structure des mots :

1. règles lexicales,
2. approche dite «mot-et-paradigme»,
3. approche structurelle ;

■ ceux qui concernent la représentation et la propagation des informations linguistiques en STTs, tels que : lequel, de la base ou du suffixe, est la tête de la construction ? Peut-on appliquer (et si oui, comment) les principes de propagation des traits tels qu'ils sont décrits en HPSG (principe des traits de tête, principe de valence, principe sémantique, schémas de dominance immédiate) ?

3.2.1. Règles lexicales

Les règles lexicales (RLs) sont proposées dans Pollard et Sag (1987 : 191-218) comme un moyen pour la construction dynamique de lexiques. En dehors de la morphologie flexionnelle, cette technique s'utilise également pour générer automatiquement les variations dans la réalisation des arguments d'un prédicat. L'exemple (20) est une application

typique de RL pour la construction des formes verbales de la 3e personne du singulier du présent indicatif, à partir des formes de base :

(20)
$$\underset{base}{\begin{bmatrix} \text{PHON } \boxed{1} \\ \text{3RDSNG } \boxed{2} \\ \text{SYN | LOC | COMPS } \boxed{3} \\ \text{SEM | CONT } \boxed{4} \end{bmatrix}} \rightarrow \underset{3rdsng}{\begin{bmatrix} \text{PHON } F_{3rdsng}(\boxed{1}, \boxed{2}) \\ \text{SYN | LOC | COMPS } \boxed{3} \\ \text{SEM | CONT } \boxed{4} \end{bmatrix}}$$

F_{3rdsng} est un opérateur qui retourne une forme verbale résultant de la concaténation de la base et du suffixe approprié. Les étiquettes (1, ...) n'indiquent pas ici de partages de valeurs internes à une STT, mais plutôt des variables à valeur identique dans les deux entrées considérées.

(20) représente donc une instruction du compilateur qui compile toute forme verbale de base en sa forme «3rdsng». On voit facilement que le lexique doit contenir un très grand nombre de RLs pour prendre en compte l'ensemble du système flexionnel. Ce mécanisme soulève par ailleurs d'autres problèmes :

1. les RLs constituent un type de donnée hétérogène dans la théorie linguistique HPSG ;
2. les RLs sont sans relation entre elles, ce qui traduit, comme le remarquent Krieger et Nerbonne (1993 : 6), une absence de généralité ;
3. le mécanisme des RLs n'est pas disponible en ALEP, de telle façon que leur simulation dans cet environnement revient à utiliser un dictionnaire de formes fléchies.

3.2.2. Approche «mot-et-paradigme»

Cette approche, introduite par Krieger et Nerbonne (1993 : 8) utilise la notion de disjonction distribuée. Ce type de disjonction rend possible la représentation d'ensembles de STTs variant en parallèle, comme l'illustre (21) qui est une adaptation en français d'un exemple allemand de Krieger et Nerbonne (1993 : 9), *i.e.* la conjugaison au présent de l'indicatif pour les verbes du premier groupe :

(21)
$$\begin{bmatrix} \text{MORPH} & \begin{bmatrix} \text{STEM } \boxed{2} \\ \text{ENDING } \boxed{3}\{\$1\ \text{'e', 'es', 'e', 'ons', 'ez', 'ent',}\} \\ \text{FORM } \boxed{2} \ \& \ \boxed{3} \end{bmatrix} \\ \text{SYN} & \begin{bmatrix} \text{LOC} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{AGR} & \left\{ \$1 \begin{bmatrix} \text{PER 1} \\ \text{NUM SG} \end{bmatrix}, \begin{bmatrix} \text{PER 2} \\ \text{NUM SG} \end{bmatrix}, \ldots \right\} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Ainsi, l'application du paradigme (21) à la base verbale `essuY` combinée au suffixe `-es` implique l'extraction de l'item $\begin{bmatrix} \text{PERS 2} \\ \text{NUM SG} \end{bmatrix}$ parmi les traits d'accord.

Les verbes irréguliers (`aller`, `être`, ...) sont eux aussi représentables au moyen de disjonctions distribuées. De plus, Bennett *et al.* (1995 : 63) montrent que cette approche s'applique aussi aux cas d'affixation multiple (`grand+e+s`).

Bien que la disjonction distribuée soit disponible dans le formalisme ALEP, nous n'adoptons pas cette approche car, une fois encore, elle constitue simplement une façon élégante pour compiler un dictionnaire de formes fléchies. Une motivation supplémen-

taire, cependant, est que toute tentative pour obtenir un certain degré de complétude rendrait les paradigmes totalement incompréhensibles et non maintenables, puisqu'un grand nombre de co-variations de STs, (éventuellement enchâssées les unes dans les autres) doivent être définies afin de refléter la complexité du système flexionnel du français. En d'autres termes, il semble que cette approche doive être réservée à des langues ayant un système flexionnel moins riche.

3.2.3. Approche structurelle

Nous avons choisi l'approche structurelle, où bases et affixes sont combinés au moyen de règles basées sur la déviation pour produire des mots. Cette approche réduit de façon optimale la taille du lexique, en définissant des entrées distinctes pour les suffixes et les bases. Ainsi, `essuient` reçoit une représentation morphosyntaxique qui est une instance de la figure (22) présentée ci-dessous.

Nous montrons ici comment, dans ce cadre, les informations linguistiques sont structurées en STs, et quels principes assurent et contrôlent la propagation de ces informations.

Dans la théorie HPSG, il existe des principes généraux pour la propagation d'informations, tels que: (a) le principe des traits de tête, selon lequel les informations de tête sont propagées récursivement depuis la branche principale (*i.e.* la tête) d'une structure, vers le père, (b) le principe de valence, qui exprime le fait que le cadre de sous-catégorisation du nœud père est identique à celui de la branche tête, à l'exception des comblements déjà combinés avec celle-ci, (c) le principe sémantique qui dit (de façon simplifiée) que la branche tête d'une structure transmet au père son contenu sémantique. Enfin, rappelons que HPSG est muni d'un nombre très limité de schémas pour la représentation des relations de dépendance: l'un d'eux, le schéma tête-complément, décrit les conditions permettant la combinaison entre un prédicat et ses compléments.

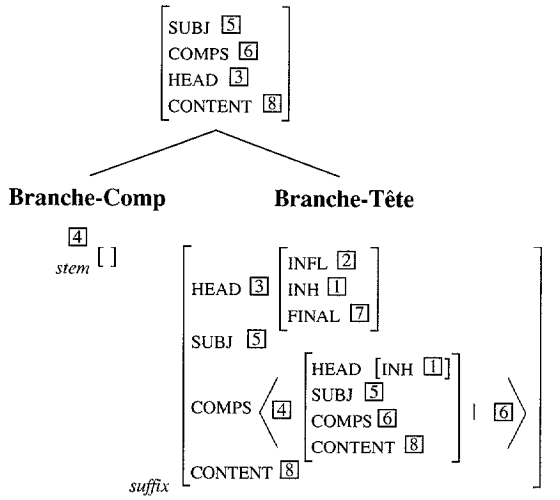
En ce qui concerne la propagation des traits morphosyntaxiques dans une structure base + suffixe, il y a trois types de décisions à prendre:

1. tête de la structure: étant donné qu'une grande partie des traits de tête morphosyntaxiques (mais pas tous, cf. section 3.3.) proviennent du suffixe, c'est celui-ci qui constitue la tête;
2. sous-catégorisation: l'affirmation ci-dessus est contredite par le fait que l'ensemble des compléments hérités par le père proviennent de la base. D'autre part, la base est elle-même sélectionnée par le suffixe. Ces deux faits ne peuvent pas, de façon immédiate, être pris en compte simultanément par le principe de valence HPSG;
3. la remarque ci-dessus vaut également pour le contenu sémantique, que le père partage avec la base, et non avec le suffixe.

Nous avons décidé de rester aussi proche que possible de HPSG, et d'en conserver l'ensemble des généralisations. Ainsi, le schéma tête-complément est appliqué aux combinaisons base-suffixe, et les principes sémantique et de valence sont maintenus. Cette décision a été rendue possible grâce à la technique d'héritage d'arguments (Hinrichs et Nakazawa 1993: 23 et Pollard 1990), et par l'emploi exclusif de règles binaires.

La figure (22) résume nos choix :

(22)



(22) montre que les traits de tête (HEAD) sont distribués selon trois attributs. INFL véhicule les informations flexionnelles depuis le suffixe (étiquette 2), INH contient les informations de tête intrinsèques de la base (étiquette 1), et FINAL est utilisé pour contrôler les conditions d'attachement récursif du suffixe.

Le suffixe sous-catégorise une base qui est le premier élément de sa liste COMPS, et hérite de tous les compléments de la base. De plus, il hérite de son contenu sémantique, de telle façon que les généralisations HPSG mentionnées ci-dessus restent vraies. Enfin, le suffixe hérite d'une partie des informations de tête de la base via la spécification d'un partage de valeur entre les traits INH (étiquette 1).

3.3. STRUCTURE DES MOTS EN FRANÇAIS

■ Intuitivement, il est clair que quelle que soit sa terminaison, une base contient toujours les mêmes informations catégorielles, le même cadre de sous-catégorisation et la même structure argumentale. Nous admettrons donc, en un premier temps, que les traits syntaxiques et sémantiques proviennent intégralement de la base : nous montrons plus loin quelles sont les limites d'une telle affirmation.

■ Tournons-nous maintenant vers les traits d'accord et de conjugaison. À première vue, il est tentant de poser comme hypothèse que la valeur des traits flexionnels d'une combinaison de morphèmes est égale à l'union des traits flexionnels de chaque morphème. Cette équation rend compte effectivement de certaines combinaisons, comme par exemple `dorm+ons`, où la base indique par défaut le présent de l'indicatif, le suffixe, la première personne du pluriel, et ces deux informations sont unifiables pour identifier la valeur flexionnelle de la combinaison. En revanche, cette équation est inappropriée dans un grand nombre de cas. Ainsi, faisons l'hypothèse que la valeur flexionnelle de `rentr+É` soit connue : participe passé **masculin** singulier[5]. L'adjonction de `e` à cette séquence de morphèmes apporte une information partiellement incompatible avec les valeurs préexistantes, *i.e.* le trait d'accord **féminin** singulier. La combinaison `rentr+É+e` entraîne donc une remise à jour de valeurs calculées précédemment.

Ces observations conduisent aux remarques suivantes :

- la grammaire utilisée pour vérifier la bonne formation des combinaisons de morphèmes doit permettre une transmission sélective des informations de chaque item vers le résultat de la combinaison ;
- de son côté, la structure du lexique doit refléter les propriétés des bases et des suffixes :
  — une base est porteuse d'informations de tête (*i.e.* sa catégorie grammaticale ainsi que les informations intrinsèques à la catégorie), qui sont propagées lors des combinaisons successives. De la même façon, le cadre de sous-catégorisation et la structure argumentale d'un mot sont hérités dans leur majorité de la base de ce mot ;
  — le temps, l'aspect et l'accord morphosyntaxique proviennent intégralement des suffixes flexionnels. Chaque suffixe transmet une information spécifique ;
  — les traits flexionnels des bases ont des valeurs instanciées par défaut : ainsi, une base verbale marque le présent de l'indicatif, une base nominale, marque le singulier, et une base adjectivale, le masculin-singulier. Ce choix est justifié par ce que l'on appelle l'affixation nulle, c'est-à-dire l'identification possible d'une base (pouvant se combiner avec un / des suffixes) avec un mot fléchi : *prend*, *enfant*, *petit* ;
  — l'adjonction de certains suffixes entraîne tout simplement l'enrichissement des informations flexionnelles de la base (`dorm`$_{presind}$ → `dorm+ons`$_{presind+1pp}$) ;
  — pour d'autres, il y a remise à jour (partielle) des informations préexistantes (`rentr`-$_{presind}$ → `rentr+É`$_{partpass+ms}$).

Ce partage des tâches (les suffixes fournissent les traits flexionnels, la base fournit le reste) a cependant une limite, qui est celle de l'indépendance entre morphologie et syntaxe. Ainsi, le français, qui est une langue fortement fléchie, impose des règles d'accord entre le verbe et son sujet[6]. Dans une approche basée sur l'unification de structures de traits typées, cet accord est reflété par un contrôle lexical de la part du verbe des traits flexionnels de son sujet. Ces traits étant imposés par le suffixe, c'est donc lui qui va déterminer les valeurs d'accord du sujet sélectionné par la base. En d'autres termes, les traits de sous-catégorisation sont fournis partiellement par la base, et partiellement par les suffixes véhiculant les informations d'accord. Cela est vrai pour l'accord sujet-verbe fini en nombre et en personne, mais également pour l'accord sujet-participe passé en genre et en nombre. Dans le dernier cas, le conflit engendré par les informations contradictoires (cf. l'exemple `rentr+É` / `rentr+É+e` ci-dessus) implique une remise à jour des informations flexionnelles contenues dans la structure décrivant le sujet.

Pour toutes les combinaisons licites base + suffixe, la règle **mot / base → base suffixe** doit :

- prendre en compte les valeurs par défaut de la base ;
- refléter l'absence de tête dans la combinaison (chacun des éléments propageant sélectivement un ensemble donné de traits) ;
- à la fois, effectuer l'unification de valeurs flexionnelles compatibles et choisir, en cas de conflit, laquelle des valeurs contradictoires est héritée par la combinaison de morphèmes.

3.4. EXEMPLES

Dans cette dernière section, nous illustrons l'approche décrite en 3.2. par l'étude détaillée de deux cas représentatifs, qui mettent en évidence les problèmes de partage de structure évoqués en conclusion de 3.3. :

1. la combinaison `rentr+i+ons+` dans laquelle se succèdent un suffixe «de mise à jour» et un suffixe d'«unification» des traits de flexion ;
2. la combinaison `rentr+É+e` pour laquelle l'accord sujet-verbe nécessite un traitement (lexical) particulier.

3.4.1. Analyse morphosyntaxique de «`rentr+i+ons`*»*

L'entrée lexicale de la base, schématisée par la matrice (23) exprime :

- ses propriétés inhérentes : c'est un verbe, dont l'auxiliaire de conjugaison est être, etc. (trait SYNT|HEAD|INH) ;
- ses valeurs flexionnelles par défaut : présent de l'indicatif, le nombre et la personne étant laissés non spécifiés (trait SYNT|HEAD|INF) ;
- son cadre de sous-catégorisation (traits SYNT|SUBJ et SYNT|COMPS), et la relation prédicat-argument correspondante (trait SEM).

(23)

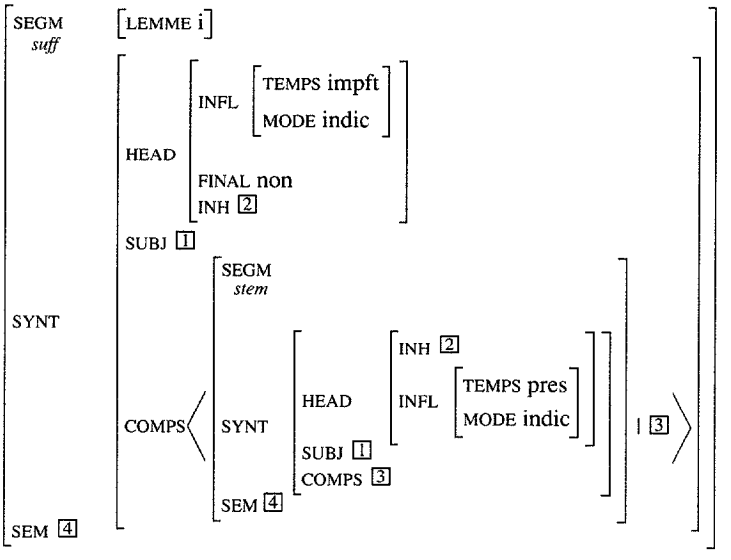
$$\begin{bmatrix} \text{SEGM} & [\text{LEMME rentr}] \\ \textit{stem} & \\ \text{SYNT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{INFL} & [\text{TEMPS pres}, \text{MODE indic}] \\ \text{INH} & [\text{CAT v}, \text{AUX etre}] \end{bmatrix} \\ \text{SUBJ} & \langle \text{NP} : \boxed{1} \rangle \\ \text{COMPS} & \langle\,\rangle \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{REL rentrer} \\ \text{ARG1 } \boxed{1} \end{bmatrix} \end{bmatrix}$$

Quant à l'emploi du suffixe `i` auquel la base se combine, il est caractérisé par les propriétés suivantes :

- le suffixe ne peut pas être employé comme fin de mot (trait SYNT|HEAD|FINAL) ;
- il sous-catégorise (trait SYNT|COMPS) une base au présent de l'indicatif, c'est-à-dire n'ayant subi au préalable aucune flexion. Cette entrée s'oppose donc à celle du suffixe marquant le conditionnel présent (`rentr+r+i+ons+`), et sélectionnant une base complexe qui indique l'indicatif futur (`rentr+r+`). Le suffixe hérite de l'ensemble des arguments de la base, ainsi que de son contenu sémantique ;
- il met à jour le temps de la base, qui permute à la valeur «imparfait», et intègre la valeur d'accord «pluriel» (trait SYNT|HEAD|INFL) qui est sous-spécifiée dans la base. La mise à jour est explicitée par l'absence de partage de valeurs entre les traits INFL du suffixe et de la base combinés.

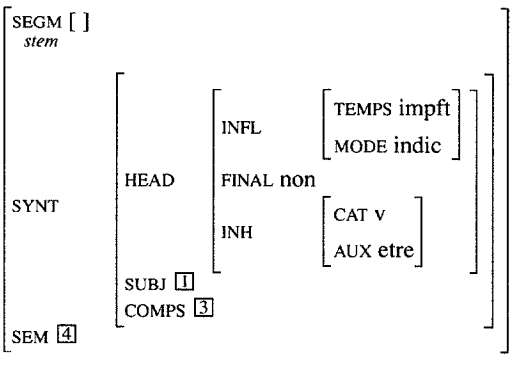
Ces propriétés apparaissent dans la figure (24)

(24)



La combinaison des deux entrées lexicales, au moyen du schéma tête-complément (cf. figure (22)), produit la structure de mot incomplète (25) (contrainte par la valeur «non» de FINAL).

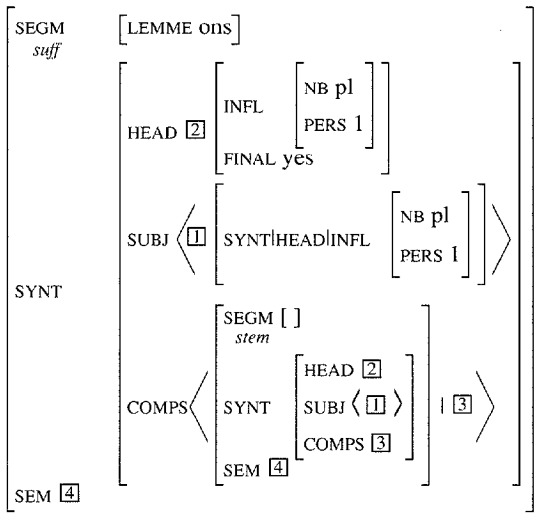
(25)



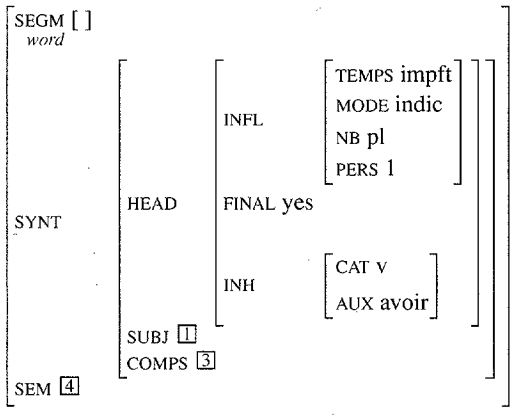
Le dernier suffixe combiné (ons) est obligatoirement le dernier morphème d'un mot ([FINAL oui]), et ses valeurs flexionnelles ne sont jamais incompatibles avec les valeurs de la base que le suffixe sélectionne, ce qui est indiqué dans la figure (26a) par un partage de valeur entre les traits INFL de ons et de la structure qui le complète. Similairement, le suffixe partage le nombre et la personne avec le sujet sous-catégorisé par la base, pour en contraindre l'accord. La structure résultante, dans (26b), est un verbe fini à la première personne du pluriel de l'indicatif imparfait et avec les valeurs syntaxiques et sémantiques appropriées :

(26)

(a)



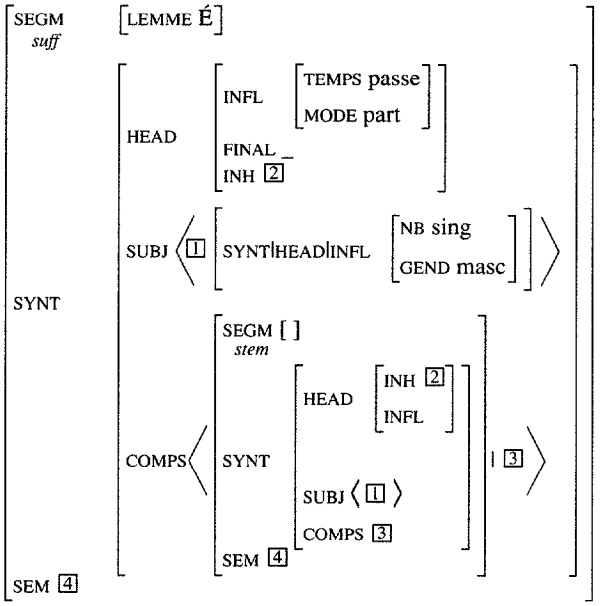
(b)



3.4.2. Analyse morphosyntaxique de « `rentr+É+e` *»*

La base (cf. (23)) est combinée au suffixe É (cf. (27)) qui indique le participe passé masculin singulier, ce qui implique une remise à jour des traits flexionnels apportés par la base, et une spécification des traits d'accord du sujet (l'accord avec un participe passé est obligatoire dès lors que l'auxiliaire de conjugaison est *être*).

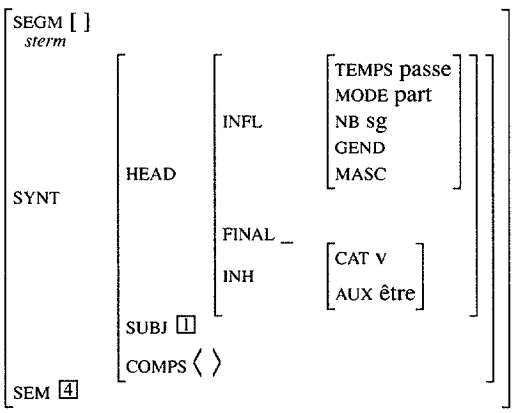
(27)



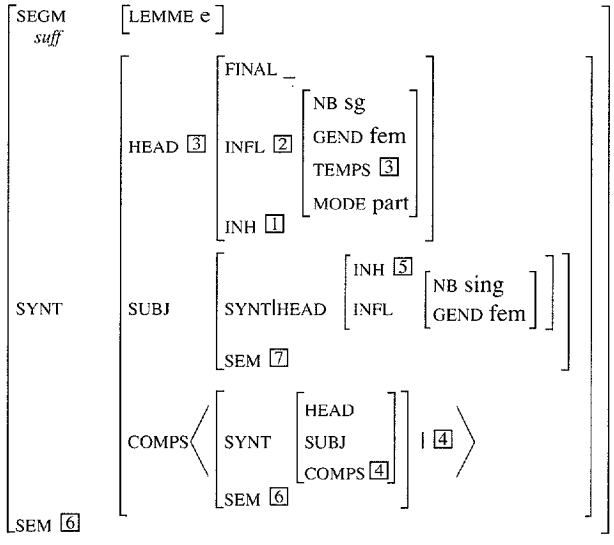
L'application du schéma tête-complément produit la structure (28a) qui peut être un mot fini (qui indique alors un participe passé masculin-singulier), ce que reflète la valeur non spécifiée du trait de tête FINAL de É et de sa projection `rentr+É`. Cette contrainte n'est pas obligatoire, ainsi la combinaison avec la marque du féminin singulier `e` est possible. L'entrée de ce suffixe, en (28b) indique la double remise à jour des traits flexionnels : les traits flexionnels du suffixe diffèrent de ceux de la base, ainsi que les traits d'accord du sujet sous-catégorisé par les deux morphèmes.

(28)

(a)

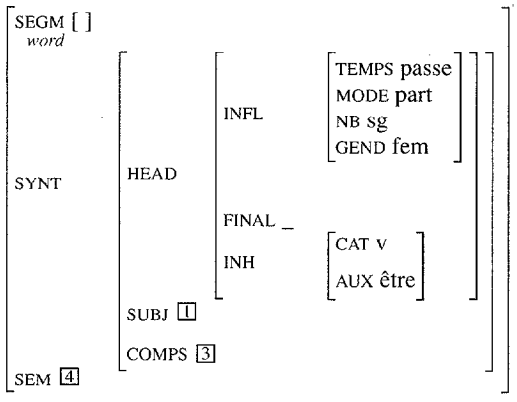


(b)



La structure résultante en (29) est un mot fini. La valeur non spécifiée de FINAL indique qu'il pourrait se combiner à un autre suffixe (plus précisément, le marqueur s du pluriel).

(29)



4. CONCLUSION

Nous avons décrit un ensemble de stratégies pour la description de la flexion du français sur la base de techniques computationnelles récentes, réalisant des concepts appartenant à HPSG, à la MDN et plus généralement au paradigme des structures de traits typées. De plus, nous avons pris en compte les problèmes d'efficacité en terme de temps d'exécution, dans la mesure où nous avons défini nos spécifications par rapport à

ALEP, qui appartient à la famille des formalismes dits «à expressivité réduite». De tels formalismes sont conçus de façon à faire de l'efficacité l'une de leurs propriétés majeures, pour permettre leur utilisation à l'extérieur d'un environnement de recherche. Bien que muni, par conséquent, d'un ensemble limité de composants formels, ALEP dispose de suffisamment d'outils pour permettre au linguiste l'expression aisée et satisfaisante de descriptions linguistiques. Nous avons en effet montré que cet ensemble restreint (en ALEP, on ne dispose pas de mécanismes pour représenter l'implication, la séparation entre dépendance immédiate et ordre linéaire, les règles lexicales, l'héritage multiple, la disjonction généralisée, les ensembles, les opérations entre listes ou ensembles et l'héritage par défaut) était assez expressif pour décrire les faits pertinents concernant la morphologie flexionnelle du français. La prochaine étape envisagée est l'extension de ce système pour produire un lexique compact et non redondant en tant que composante d'une grammaire «noyau» automatique efficace pour l'analyse et la synthèse du français.

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV^es^ Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

1. Cette présentation rapporte des résultats de travaux financés par la CEE dans le cadre des programmes MLAP (Multi-Lingual-Action-Plan) et LRE (Linguistic Research and Engineering).
2. Ce symbole est généré automatiquement lors de la phase préalable de préparation du document à analyser.
3. L'étoile indique l'agrammaticalité.
4. Nous nous plaçons ici du point de vue de l'analyse. La démarche inverse en génération, à savoir la répartition des traits entre morphèmes, est un processus immédiat dans un formalisme basé sur l'unification.
5. On fait l'hypothèse d'une lecture intransitive du verbe.
6. Nous laissons de côté l'accord entre le verbe et son objet direct, car il sous-entend une variation dans la réalisation syntaxique de l'objet — qui doit être clitique ou relatif — et donc dépasse le cadre de la morpho-syntaxe. La représentation lexicale d'un tel accord doit plutôt se voir comme le résultat de l'application de règles de redondance lexicales (cf. Pollard et Sag 1987 : 191-218).

RÉFÉRENCES

ALSHAWI, H., ARNOLD, D.-J., BACKOFEN, R., CARTER, D.-M., LINDOP, J., NETTER, K., PULMAN, S.-G., TSUJII, J. et H. USZKOREIT (1991) : *Eurotra ET/61: Rule Formalism and Virtual Machine Design Study (Final Report)*, Luxembourg, CEE.

BENNETT, P., SCHMIDT, P. et A. THEOFILIDIS (1995) : «Deliverable D4 — WP4: Morphology», *MLAP93-15*, Luxembourg, CEE.

BESCHERELLE, M. (1990) : *L'art de conjuguer*, Paris, Hatier.

BLACK, A., RITCHIE, G., PULMAN, S. et G. RUSSELL (1987) : «Formalisms for Morphographemic Description», *EACL-3*.

HINRICHS, E. et T. NAKAZAWA (1993) : «Aspects of German VP Structure», *An IIPSG Account*, «SfS-Report-01-93», Tübingen.

KOSKIENNEMI, K. (1983) : «Two-level Model for Morphological Analysis», *Proceedings of 8th IJCAI*, Karlsruhe, pp. 683-685.

KRIEGER, H.-U. (1994) : «Derivation without Lexical Rules», C. Rupp *et al.* (Eds), *Constraints, Language and Computation*, Academic Press, pp. 277-313.

KRIEGER, H.-U. et J. NERBONNE (1993) : «Feature-Based Inheritance Networks for Computational Lexicons», *Research Report 31*, Saarbrüken, Deutsches Forschungszentrum für Künstliche Intelligenz.

POLLARD, C. (1990) : «On Head-Non Movement», *Proceedings of the Symposium on Discontinuous Constituency*, Tilburg.

POLLARD, C. et I. SAG (1987) : *Information-Based Syntax and Semantics*, vol. 1, CSLI.

POLLARD, C. et I. SAG (1994) : *Head-Driven Phrase Structure Grammar*, CSLI / University of Chicago Press.

RITCHIE, G., RUSSELL, G., BLACK, A. et S. PULMAN (1992) : *Computational Morphology*, MIT Press.

RUESSINK, H.-A. (1990) : «Two-Level Formalisms», P. Coopmans, B. Schouten, W. Zonneveld (Eds), *OTS Yearbook, University of Utrecht*.

TROST, H. (1990) : «The Application of Two-Level Morphology to Non-concatenative German Morphology», *Proceedings of COLING-90*, vol. 2, Helsinki, pp. 371-376.