Apprentissage automatique d'un chunker pour le français

Isabelle Tellier[1,2], Denys Duchier[2], Iris Eshkol[3],
Arnaud Courmet[2], Mathieu Martinet[3]
(1) LaTTiCe, université Paris 3 - Sorbonne Nouvelle
(2) LIFO, université d'Orléans
(3) LLL, université d'Orléans
isabelle.tellier@univ-paris3.fr, denys.duchier@univ-orleans.fr,
iris.eshkol@univ-orleans.fr, arnaud.coumet@gmail.com,
mathieu_martinet@hotmail.fr

RÉSUMÉ
Nous décrivons dans cet article comment nous avons procédé pour apprendre automatiquement un chunker à partir du French Tree Bank, en utilisant les CRF (Conditional Random Fields). Nous avons réalisé diverses expériences, pour reconnaître soit l'ensemble de tous les chunks possibles, soit les seuls groupes nominaux simples. Nous évaluons le chunker obtenu aussi bien de manière interne (sur le French Tree Bank lui-même) qu'externe (sur un corpus distinct transcrit de l'oral), afin de mesurer sa robustesse.

ABSTRACT
Machine Learning of a chunker for French

We describe in this paper how to automatically learn a chunker for French, from the French Tree Bank and CRFs (Conditional Random Fields). We did several experiments, either to recognize every possible kind of chunks, or to focus on simple nominal phrases only. We evaluate the obtained chunker on internal data (i.e. also extracted from the French Tree Bank) as well as on external (i.e from a distinct corpus) ones, to measure its robustness.

MOTS-CLÉS : chunking, apprentissage automatique, French Tree Bank, CRF.

KEYWORDS: chunking, Machine Learning, French Tree Bank, CRF.

1 Introduction

Nous présentons dans cet article la démarche ayant permis d'apprendre automatiquement à partir du French Tree Bank (Abeillé *et al.*, 2003) un "chunker" ou analyseur syntaxique superficiel (Abney, 1991) du français. Alors que cette tâche a fait l'objet du challenge CoNLL 2000 [1] pour l'anglais, aucun chunker du français n'avait encore, semble-t-il, été appris automatiquement à partir de données annotées. Ce travail est une suite naturelle à l'acquisition d'un étiqueteur morpho-syntaxique (ou POS) du français réalisée précédemment à partir du même corpus

1. http ://www.cnts.ua.ac.be/conll2000/chunking/

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 431–438,
Grenoble, 4 au 8 juin 2012.

(Constant *et al.*, 2011), et sur lequel il s'appuie. Comme l'étiqueteur POS, notre chunker a été appris à l'aide des CRF (Conditional Random Fields) (Lafferty *et al.*, 2001; Tellier et Tommasi, 2011). Comme lui, il est librement disponible en téléchargement (Tellier *et al.*, 2012).

La notion de chunk peut recouvrir plusieurs niveaux de détails possibles, suivant que l'on se concentre sur les seuls groupes nominaux simples, à la façon de (Sha et Pereira, 2003) ou sur l'ensemble de tous les constituants non récursifs possibles. Ces deux variantes ont été testées et évaluées par validation croisée sur le corpus, en s'appuyant sur un étiquetage POS parfait, et montrent qu'identifier les différents types de chunks est de difficulté variable suivant leur nature. De plus, la variante qui se concentre sur les groupes nominaux simples a aussi été testée sur un autre corpus totalement différent, constitué de transcriptions de l'oral et annoté avec notre étiqueteur, donc imparfaitement. Ces évaluations permettent de mesurer la sensibilité du modèle à des conditions d'utilisation dégradées.

L'article suit la structure suivante. Tout d'abord, nous évoquons la tâche de chunking et ses différentes variantes. Nous décrivons ensuite les différentes instances du French Tree Bank de Paris 7 qui ont permis la constitution du corpus d'apprentissage ainsi que les CRF qui ont été utilisés pour cet apprentissage. Nous fournissons enfin les résultats de nos différentes expériences.

2 Le chunking

2.1 Le chunking du français

Les chunks sont des constituants continus et non-récursifs (Abney, 1991). Ils définissent la structure syntaxique superficielle des phrases et, à ce titre, sont moins coûteux et plus faciles à obtenir que leur structure en constituants complète. Pour certains textes non normés (transcriptions de l'oral par exemple), ils représentent le degré d'analyse le plus poussé qu'on puisse espérer.

A notre connaissance, peu de solutions spécifiques sont disponibles pour le chunking du français, et celles qui existent ont été écrites à la main :
– soit pour réaliser une analyse syntaxique superficielle de textes non normés, en particulier ceux transcrits de l'oral (Antoine *et al.*, 2008; Blanc *et al.*, 2010)
– soit en tant que composant d'un analyseur syntaxique complet, comme par exemple les systèmes ayant participé aux campagnes d'évaluation Easy et Passage (Paroubek *et al.*, 2006)
– soit encore en tant que composant d'une plateforme généraliste et multilingue comme Gate [2]
Nous proposons à la place de coder la tâche de chunking comme une annotation, et de l'apprendre automatiquement à l'aide d'un CRF, en nous inspirant des expériences de (Sha et Pereira, 2003).

2.2 Découpages en chunks

La notion de chunk n'est pas toujours très précisément définie. Deux niveaux de détails sont possibles pour caractériser les chunks :

2. http ://www.semanticsoftware.info/munpex

– soit on s'intéresse aux seuls groupes nominaux simples (i.e. non récursifs), qui sont chacun constitués d'un unique nom ou pronon, incluant ses éventuels groupes adjectivaux immédiats, déterminants et adjectifs numériques. Les compléments du nom sont dans des chunks distincts de celui du nom qu'ils qualifient.
– soit on s'intéresse à tous les groupes possibles, en cherchant à obtenir un parenthésage complet de la phrase. Dans ce cas, les différents types possibles de chunks, tels qu'ils apparaissent dans le French Tree Bank, sont :
  – les groupes nominaux ou NP, définis comme précédemment sauf quand ils sont inclus dans un des autres types suivants ;
  – les groupes verbaux ou VN, incluant les formes interrogatives, infinitives, modales.. ;
  – les groupes prépositionnels ou PP, incluant tous les groupes nominaux introduits par une préposition ainsi que tous ceux qui qualifient les VN ;
  – les groupes adjectivaux ou AP, incluant les éventuels adverbes modifieurs d'adjectifs ;
  – les groupes adverbiaux ou Adp, incluant les modifieurs de phrases ;
  – les groupes coordonnés ou COORD, introduits par une conjonction de coordination, et pouvant aussi inclure des groupes nominaux.

Ces différents chunks peuvent bien sûr être obtenus à partir de la structure en constituants de la phrase. Par exemple, l'arbre de la Figure 1 donne lieu aux deux découpages suivants :
– (La commercialisation efficace)$_{NP}$ est plus exigeante.
– (La commercialisation efficace)$_{NP}$ (est)$_{VN}$ (plus exigeante)$_{AP}$.

Dans le cas des compléments du nom ou des groupes nominaux coordonnés par exemple, le découpage de premier type n'est pas strictement inclu dans celui de deuxième type, comme l'illustre le cas suivant :
– (La commercialisation)$_{NP}$ de (la marchandise)$_{NP}$ et (des services)$_{NP}$ est plus exigeante.
– (La commercialisation)$_{NP}$ (de la marchandise)$_{PP}$ (et des services)$_{COOR}$ (est)$_{VN}$ (plus exigeante)$_{AP}$.

Pour aborder la tâche de chunking comme une tâche d'annotation, il suffit d'associer à chaque mot appartenant à un chunk une étiquette donnant son type (voit NP, soit un type parmi {NP, VN, PP, AP, AdP, VCOOR}) accompagnée du codage BIO (Begin/In/out) qui permet de délimiter ses frontières. Dans le cas d'un parenthésage total, le type O est inutile car la fin d'un chunk coincide toujours avec le début d'un autre :
– La/B-NP commercialisation/I-NP efficace/I-NP est/O plus/O exigeante/O.
– La/B-NP commercialisation/I-NP efficace/I-NP est/B-VN plus/B-AP exigeante/I-AP.

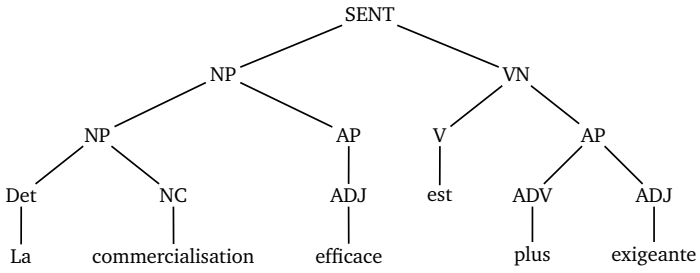

FIGURE 1 – Arbre d'analyse syntaxique extrait du French Tree Bank

3 Constitution de corpus de référence

3.1 Le French Tree Bank

Le French Tree Bank (Abeillé *et al.*, 2003) est la ressource à partir de laquelle nous avons pu constituer des ensembles d'exemples de référence pour l'apprentissage automatique. Ce corpus est composé d'articles du journal "Le Monde". Les phrases qui y figurent sont complètement analysées syntaxiquement en constituants, comme dans la Figure 1. Nous en avons extrait deux variantes de corpus annotés en chunks, correspondant aux deux notions possibles de chunks évoquées précédemment.

3.2 Homogénéisation des étiquettes

Pour ce travail, nous disposions en fait de deux versions complémentaires du corpus :
– la version arborée, composée d'environ dix mille fichiers XML (un par phrase). Ces fichiers décrivent donc la structure syntaxique complète des phrases ainsi que de leurs unités. Les mots sont associés à une liste d'attributs qui les caractérisent (lemme, catégorie, ...).
– une version où ne figurent plus que les mots et leur catégorie morpho-syntaxique, ayant servi à acquérir un étiqueteur (Constant *et al.*, 2011). Son jeu d'étiquettes comprend 29 catégories POS distinctes. Ces catégories ne correspondent pas exactement avec la valeur de l'attribut "cat" associé aux mots de la version arborée (des simplifications ont eu lieu), ce qui nous a containt à quelques prétraitements.

Il était indispenable d'harmoniser les catégories morpho-syntaxiques figurant dans ces deux versions du corpus, car notre chunker doit pouvoir s'appuyer sur l'étiqueteur POS appris précédemment à partir des catégories simplifiées. L'étiqueteur POS utilisé ne prend pas en compte pour l'instant la reconnaissance des unités multimots.

4 Le modèle d'apprentissage

4.1 Les CRF

Les champs markoviens conditionnels ou CRF (Tellier et Tommasi, 2011) sont des modèles probabilistes discriminants introduits par (Lafferty *et al.*, 2001) pour l'annotation de données séquentielles. Ils ont été utilisés dans de nombreuses tâches de traitement automatique des langues, pour lesquelles ils sont particulièrement bien adaptés (McCallum et Li, 2003; Sha et Pereira, 2003; Tsuruoka *et al.*, 2009; Tellier *et al.*, 2010).

Les CRF permettent d'associer à une observation x une annotation y en se basant sur un ensemble d'exemples annotés (x, y). La plupart du temps (et ce sera le cas ici), x est une *séquence d'unités* (ici, une suite d'unités lexicales associées à leur catégorie POS) et y la *séquence des annotations correspondante* (ici, la suite des étiquettes BIO couplées au type des chunks). Ils sont définis par X et Y, deux champs aléatoires décrivant respectivement chaque unité de l'observation x et de son annotation y, et par un graphe dont $V = X \cup Y$ est l'ensemble des nœuds (vertices) et $E \subseteq V \times V$ l'ensemble des arcs (edges). Deux variables sont reliées dans le graphe

si elles dépendent l'une de l'autre. Le graphe sur le champ Y des CRF linéaires est une simple chaîne qui traduit le fait que chaque étiquette est supposée dépendre de l'étiquette précédente et de la suivante et, implicitement, de la donnée x complète.

Dans un CRF linéaire, on a la relation suivante :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathscr{C}} \exp\Big(\sum_k \lambda_k f_k(y_i, x, c) \Big) \quad avec$$

- $Z(x)$ est un coefficient de normalisation, défini de telle sorte que la somme sur y de toutes les probabilités $p(y|x)$ pour une donnée x fixée soit égale à 1.
- $\mathscr{C}$ est l'ensemble des cliques (sous-graphes complètement connectés) de $\mathscr{G}$ sur Y : dans le cas des CRF linéaires, ces cliques sur Y sont constituées soit d'un nœud isolé, soit d'un couple de nœuds successifs.
- Les fonctions f_k sont appelées *fonctions caractéristiques* (features) : elles sont définies à l'intérieur de chaque clique c et sont à valeurs réelles, mais souvent choisies pour donner un résultat binaire (0 ou 1). Elles doivent être founies au système par l'utilisateur. Par définition, la valeur de ces fonctions peut dépendre des annotations y_c présentes dans une certaine clique c ainsi que de la valeur de x *n'importe où dans la donnée* (et pas uniquement aux indices correspondants à la clique c, ce qui donne beaucoup d'expressivité aux CRF).
- y_c est l'ensemble des valeurs prises par les variables de Y sur la clique c pour une annotation y donnée : ici, c'est donc soit la valeur y_i d'une seule étiquette soit celles d'un couple d'étiquettes successives (y_i, y_{i+1}).
- Les poids λ_k, à valeurs réelles, permettent d'accorder plus ou moins d'importance à chaque fonction f_k dont ils caractérisent le *pouvoir discriminant*. Ce sont les paramètres du modèle : l'enjeu de la phase d'apprentissage est de fixer leur valeur en cherchant à maximiser la log-vraisemblance sur l'ensemble des exemples annotés constituant le corpus d'apprentissage.

Le logiciel que nous avons utilisé est Wapiti avec pénalisation L1 (Lavergne *et al.*, 2010), reconnu comme actuellement le plus efficace pour les CRF linéaires, car il procède lors de sa phase d'apprentissage à une *sélection* des features les plus discriminantes.

4.2 Features utilisées

Pour nos expériences, nous nous sommes contenté de features simples. Le seul attribut associé aux mots M est leur catégorie POS, notée C. A chaque fois qu'un couple $x_i = (M_i, C_i)$ est associé à une étiquette $y_i = E_i$ à une position i dans le corpus d'apprentissage, on crée une feature "unigramme" (c'est-à-dire ne prenant en compte qu'une seule étiquette) caractérisant l'association du mot et de l'étiquette, ainsi qu'une autre caractérisant l'association de la catégorie et de l'étiquette. On fait de même avec chacun des mots situés dans une fenêtre de taille 5 (de deux places avant à deux places après) centrée sur le mot courant. Les features "bigrammes" (c'est-à-dire portant sur un couple d'étiquettes successives) sont construites de la même façon, en ne tenant compte que des catégories et pas des mots, parce qu'elles varient moins que ces derniers. Les features sont donc toutes les configurations attestées dans les exemples de la forme suivante :

- $f_{1,i,j}(y_i, x) = 1$ si $y_i = E_i$ et $mot_j = M_j$, $\forall j \in [i-2, i+2]$ (=0 sinon)
- $f'_{1,i,j}(y_i, x) = 1$ si $y_i = E_i$ et $POS_j = C_j$, $\forall j \in [i-2, i+2]$ (=0 sinon)
- $f_{2,i,j}(y_i, y_{i+1}, x) = 1$ si $y_i = E_i$ et $y_{i+1} = E_{i+1}$ et $POS_j = C_j$, $\forall j \in [i-2, i+2]$ (=0 sinon)

5 Les résultats

5.1 Validation interne

Les premières évaluations ont été réalisés par validation croisée en répartissant le corpus d'apprentissage dans 5 ensembles distincts, 4 servant pour l'apprentissage et 1 pour le test. Dans chacun de ces ensembles, on dispose d'un étiquetage POS parfait, puisqu'il est lui-même issu du French Tree Bank. Un chunk est considéré comme reconnu si à la fois ses frontières et son type sont corrects.

Les seuls "groupes nominaux simples" NP sont identifiés avec une précision de 97,49%, un rappel de 97,40% et une F-mesure de 97,45. Ces excellents résultats dépassent ceux obtenus pour la tâche CoNLL 2000 sur l'anglais (où les meilleurs dépassaient à peine 94 points de F-mesure), mais ces comparaisons sont à prendre avec précautions, car ni les données ni le jeu de catégories POS utilisées n'étaient les mêmes.

Il faut remarquer que, dans le cas du chunking complet, une erreur de frontière rend erronés les deux chunks que cette frontière devrait séparer. Les taux d'erreurs sont donc naturellement globalement plus bas :

type de chunk	proportion (%)	Précision (%)	Rappel (%)	F-mesure
AP	10	68,36	68,61	68,49
AdP	2	53,57	39,47	45,45
COORD	6	80,81	76,35	78,52
NP	26	84,99	86,10	85,54
PP	34	77,79	77,82	77,81
VN	22	83,3	85,52	84,39

Ce tableau montre que les NP sont les mieux reconnus, mais avec tout de même près de 12 points de F-mesure de moins que quand ils sont la seule cible, les goupes adverbiaux étant quant à eux à la fois les plus rares et les plus difficiles à identifier. La "micro-average" (moyenne des F-mesure pondérées par les effectifs des différents chunks) vaut 79,73, tandis que la "macro-average" (moyenne donnant autant d'importance à chaque type de chunk, indépendamment de sa fréquence d'apparition) vaut : 73,37. Il n'y a pourtant pas toujours corrélation entre la fréquence d'un chunk et sa propention à être reconnu. Ainsi, PP est le type de chunks le plus fréquent car il couvre à la fois les compléments du nom qui suivent un NP et les groupes prépositionnels associés à un VN. Cette variabilité de construction explique sans doute la relative difficulté à les retrouver. Inversement, les COORD sont assez rares, mais comme ils doivent être nécessairement introduits par une conjonction de coordination, ils ne sont pas si durs à repérer.

Les résultats de notre système de chunking complet sont moins bons que ceux habituellement obtenus par les analyseurs syntaxiques complets (qui peuvent atteindre une exactitude d'environ 85%) : la simple identification des chunks est apparemment plus difficile quand elle n'est pas couplée avec celle des relations qu'ils entretiennent les uns avec les autres.

Nous aurions pu mesurer l'importance de la catégorie POS sur cette identification en cherchant à retrouver les chunks à partir d'un corpus annoté par notre étiqueteur, c'est-à-dire imparfaitement. Cependant, cet étiqueteur POS a été appris sur ce même corpus, il y fait moins de 2 points d'erreur en exactitude (puisqu'il n'en faisait déjà pas beaucoup plus en validation croisée), et l'ef-

fet de ces très rares erreurs sur le chunking sera donc difficile à mesurer. A la place, nous avons testé le résultat final du traitement : étiquetage + chunking NP sur un corpus complètement différent.

5.2 Test sur un corpus oral

Afin de tester la robustesse du chunker qui se concentre sur les groupes nominaux simples dans un contexte différent de celui dans lequel il a été appris, nous avons évalué ses performances sur un extrait du corpus de transcriptions orales ESLO [3]. Le corpus a été annoté en POS avec SEM [4], l'étiqueteur POS appris sur le French Tree Bank, sans que les catégories fournies par ce programme ne soient corrigées. Seuls les résultats du chunking ont, eux, été vérifiés à la main. Sur un corpus comprenant 575 "phrases" (i.e. tours de parole ou "groupe de souffle") et environ 9 280 mots, la performance de notre chunker tombe à moins de 40 en F-mesure, très loin de ses 97,45 points obtenus par validation croisée. L'exactitude de l'étiquetage B_NP est d'environ 56%, celui des I_NP de 61%.

Il n'est pas facile d'analyser la raison de ces résultats. Certaines erreurs semblent provenir de la segmentation qui n'est pas traitée par notre étiqueteur POS : les mots composés, entités nommées ou expressions figées devraient rester dans le même chunk et ne pas être considérés comme des compléments du nom. Les irrégularités propres à l'oral (disfluences, hésitations, amorces) sont aussi courantes et rendent bien sûr l'étiquetage POS moins fiable (même si nous n'avons pas mesuré la qualité de l'étiquetage POS indépendamment de celle du chunking), donc la reconnaisance des chunks plus délicate. En fait, la notion même de chunk doit être amendée dans ce contexte. En effet, quand le nom principal d'un chunk est oralement répété, les deux formes transcrites sont incluses dans le même chunk qui comporte donc deux noms, ce qui est en principe interdit par notre définition des NP. Si la répétition d'un déterminant ne provoque pas un changement de chunk NP, en revanche celle d'un pronom en entraîne un : est-ce toujours souhaitable ? Et doit-on considérer que des interruptions comme "heu", "oui", "ah bon" doivent être incluses dans le chunk NP qui les englobe, le découper en deux NP distincts ou en constituer un nouveau à part ? Le statut syntaxique de ces formes propres à l'oral reste sujet à discussion.

6 conclusion

Dans cet article, nous avons présenté comment obtenir efficacement deux variantes de chunkers du français par apprentissage automatique à partir du French Tree Bank.

Nos expériences montrent que la tâche de chunking est de difficulté très variable en fonction du contexte dans lequel on l'applique. La reconnaissance des NP seuls dans des textes normés ne pose pas de problèmes, mais ils sont difficiles à distinguer des autres groupes qui peuvent aussi intégrer des noms dans le cas d'un chunking complet. Enfin, la robustesse d'un chunker acquis par apprentissage automatique est très limitée quand on l'applique à des types de textes présentant des propriétés très différentes. La notion même de chunking est peut-être à préciser dans le cas des corpus oraux.

3. http ://eslo.in2p3.fr
4. http ://www.lattice.cnrs.fr/sites/itellier/SEM.html

Il nous reste à étudier en quoi la reconnaissance des unités multimots dans la phase préliminaire d'étiquetage modifie ou non les propriétés du chunking, et à repérer les dépendances entre chunks, pour se rapprocher des performances des analyseurs syntaxiques profonds. Il est aussi envisageable d'apprendre directement un segmenteur-étiqueteur POS-chunker en une seule étape, afin d'éviter de cumuler les erreurs.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Abney, S. (1991). Parsing by chunks. *In* Berwick, R., Abney, R. et Tenny, C., éditeurs : *Principle-based Parsing*. Kluwer Academic Publisher.

Antoine, J.-Y., Mokrane, A. et Friburger, N. (2008). Automatic rich annotation of large corpus of conversational transcribed speech : the chunking task of the epac project. *In Proceedings of LREC'2008*.

Blanc, O., Constant, M., Dister, A. et Watrin, P. (2010). Partial parsing of spontaneous spoken french. *In Proceedings of LREC'2010*.

Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A. et Billot, S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de TALN'11*.

Lafferty, J., McCallum, A. et Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of ICML 2001*, pages 282–289.

Lavergne, T., Cappé, O. et Yvon, F. (2010). Practical very large scale CRFs. *In Proceedings of ACL'2010*, pages 504–513. Association for Computational Linguistics.

McCallum, A. et Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *In Proceedings of CoNLL'03*.

Paroubek, P., Robba, I., Vilnat, A. et C., A. (2006). Data annotations and measures in easy, the evaluation campain for parsers of french. *In Proceedings of LREC'2006*, pages 315–320.

Sha, F. et Pereira, F. (2003). Shallow parsing with conditional random fields. *In Proceedings of HLT-NAACL 2003*, pages 213 – 220.

Tellier, I., Dupont, Y. et Courmet, A. (2012). Un segmenteur-étiqueteur et un chunker pour le français. *In Actes de TALN'12, session démo*.

Tellier, I., Eshkol, I., Taalab, S. et Prost, J. P. (2010). Pos-tagging for oral texts with crf and category decomposition. *Research in Computing Science*, 46:79–90. Special issue "Natural Language Processing and its Applications".

Tellier, I. et Tommasi, M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. *In* Eric Gaussier et François Yvon, éditeurs : *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès.

Tsuruoka, Y., Tsujii, J. et Ananiadou, S. (2009). Fast full parsing by linear-chain conditional random fields. *In Proceedings of EACL 2009*, pages 790–798.