Validation d'une expertise textuelle basée sur l'intensité d'implication

Jérôme DAVID, Fabrice GUILLET, Vincent PHILIPPE,
Henri BRIAND, Régis GRAS

LINA – École Polytechnique de l'université de Nantes
PerformanSe SAS

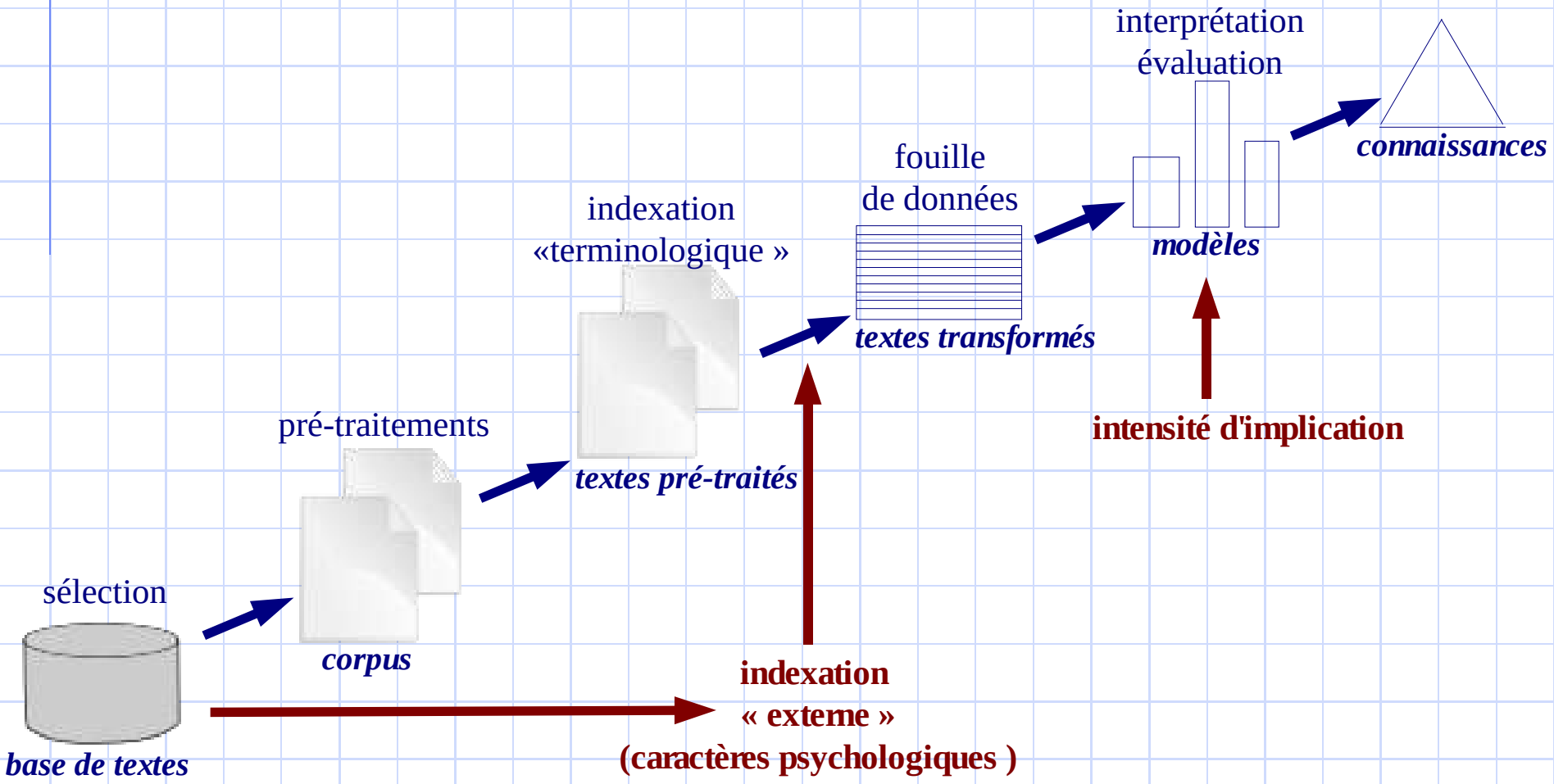
Atelier DKQ – EGC 2005

Introduction

- Objectif :
  - Rapprocher des termes par rapport à des caractères psychologiques.
  - Règles d'association du type **terme → caractère**
- Objectif de l'expert
  - Évaluer le vocabulaire exprimant les caractères.

Introduction

Processus de fouille de texte



Plan

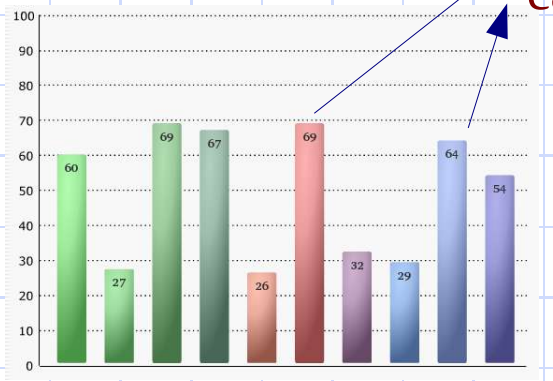
- Présentation des données
- Méthodologie
  - extraction de termes
  - modélisation des données.
  - évaluation des règles : terme -> caractère par intensité d'implication
  - formation des groupes de termes
- Résultats
  - démarche de validation des groupes de termes.
  - études des regroupements
- Conclusion / perspectives

Présentation des données

Le logiciel PERFORMANSE-Dialecho

questionnaire

calcul des scores



caractères

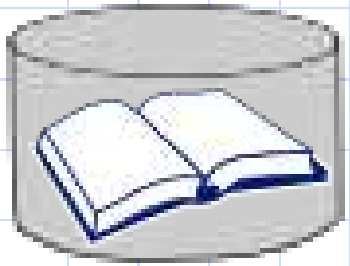
E
P
N
AFL
...

histogramme
profil psychologique

rapport

termes

génération
du rapport

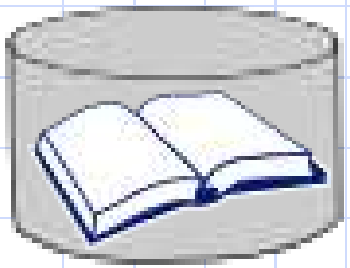


base de textes structurée
par des combinaisons de
caractères

Présentation des données

- Base textuelle structurée par des combinaison de caractères :
  - exemple :
  - SI « E0 et P0 et N0 et AFL+ et ... » ALORS

« Vous êtes d'emblée perçu comme une personne agréable et très mesurée ... »

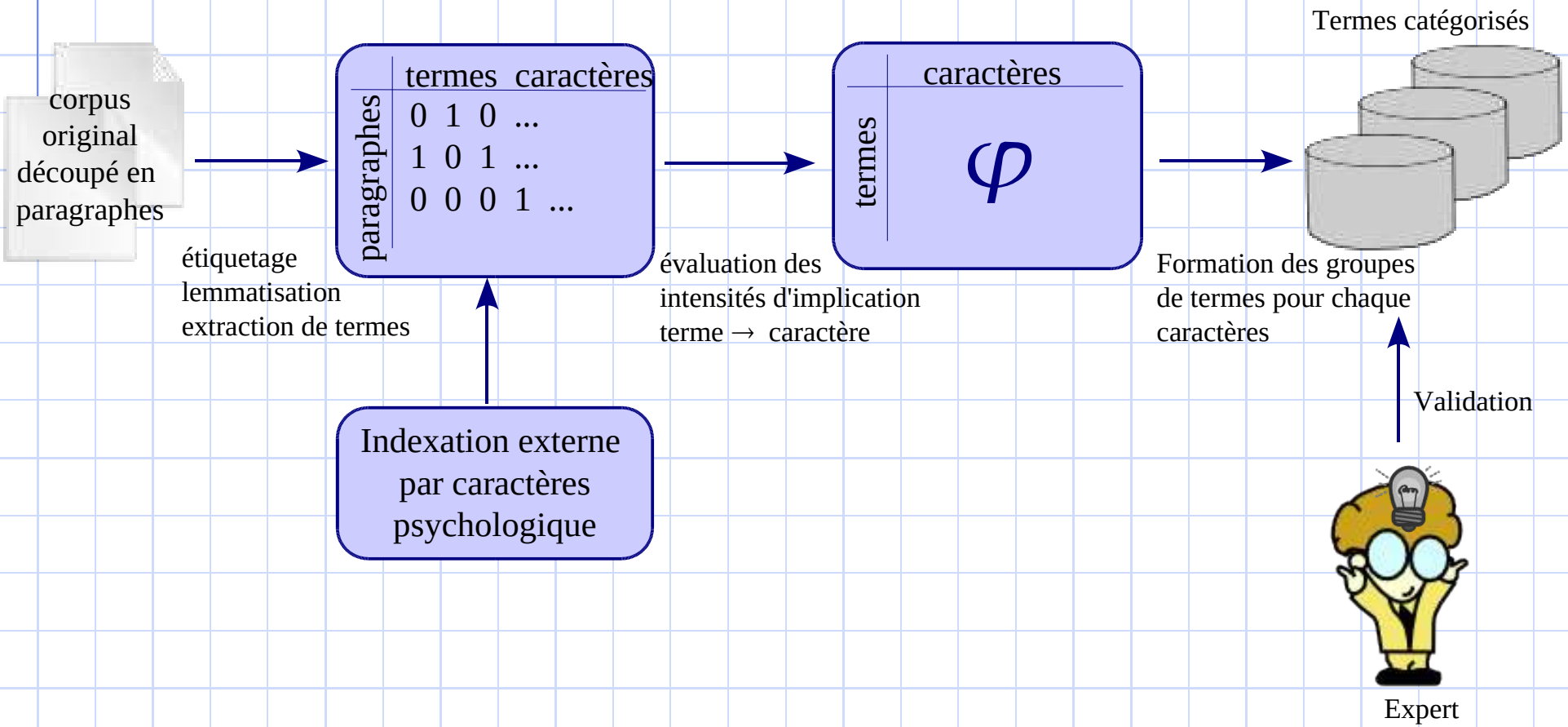


- environ 2500 pages de textes

- 12000 paragraphes (ou documents)

- 30 traits de caractères (3 positionnements possibles par dimension)

Méthodologie suivie



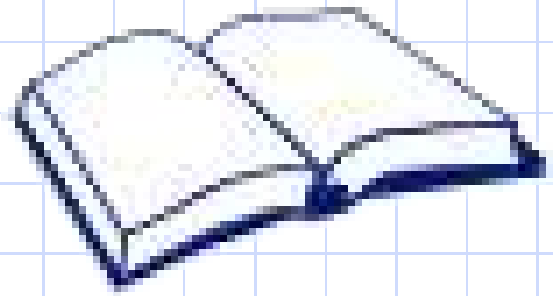
Extraction de termes

- Consiste à modéliser du texte par un ensemble de variables qui sont les termes qu'il contient

Vous êtes d'emblée perçu comme une personne agréable et très mesuré dans ses propos comme dans ses actes, et il est facile de s'entendre avec vous, car vous cherchez vous-même à entretenir un esprit d'équipe. Après un premier contact, on peut assurément estimer que vous manifestez une volonté d'adaptation qui se fonde, notamment, sur votre équilibre et sur votre tendance à relativiser les événements et à éviter les prises de position extrêmes.

→

personne agréable
personne mesurée
esprit d'équipe
premier contact
volonté d'adaptation
prise de position

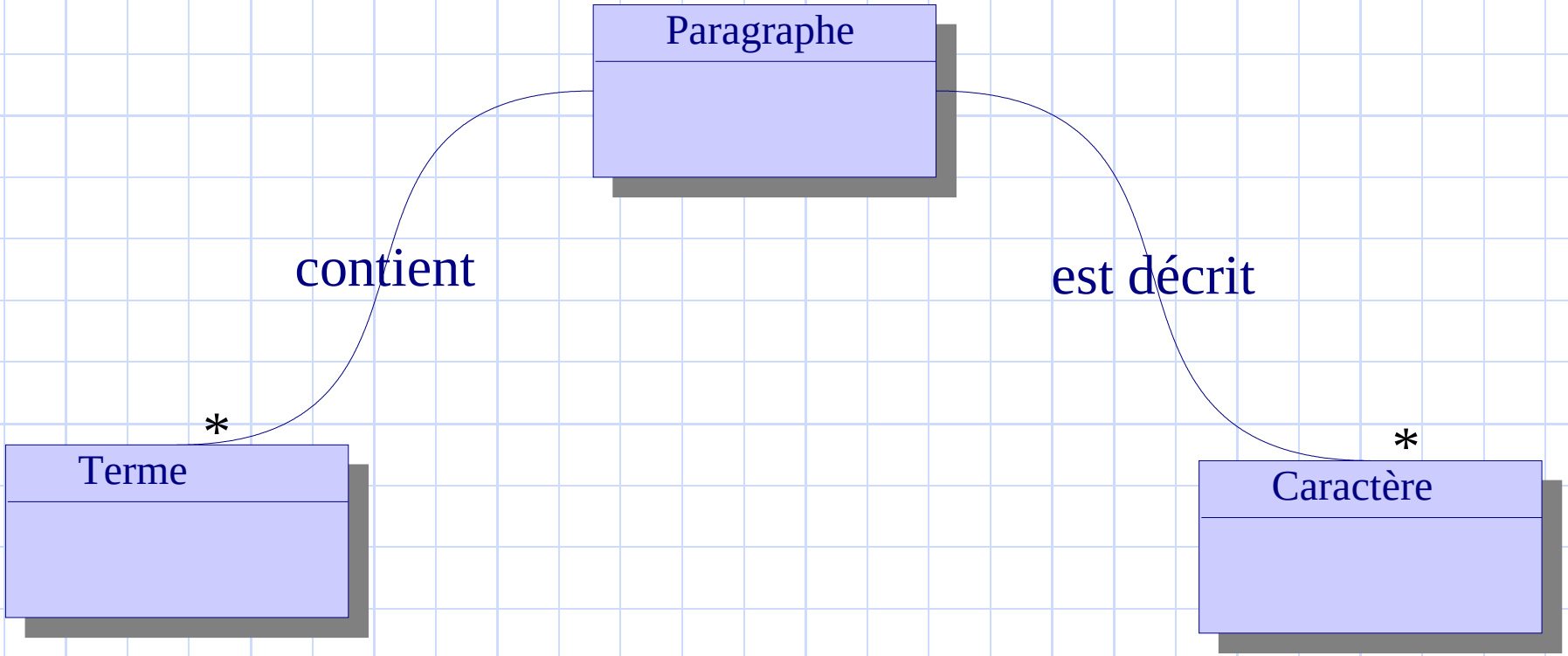


→

	termes
documents	1 si terme présent dans doc 0 sinon

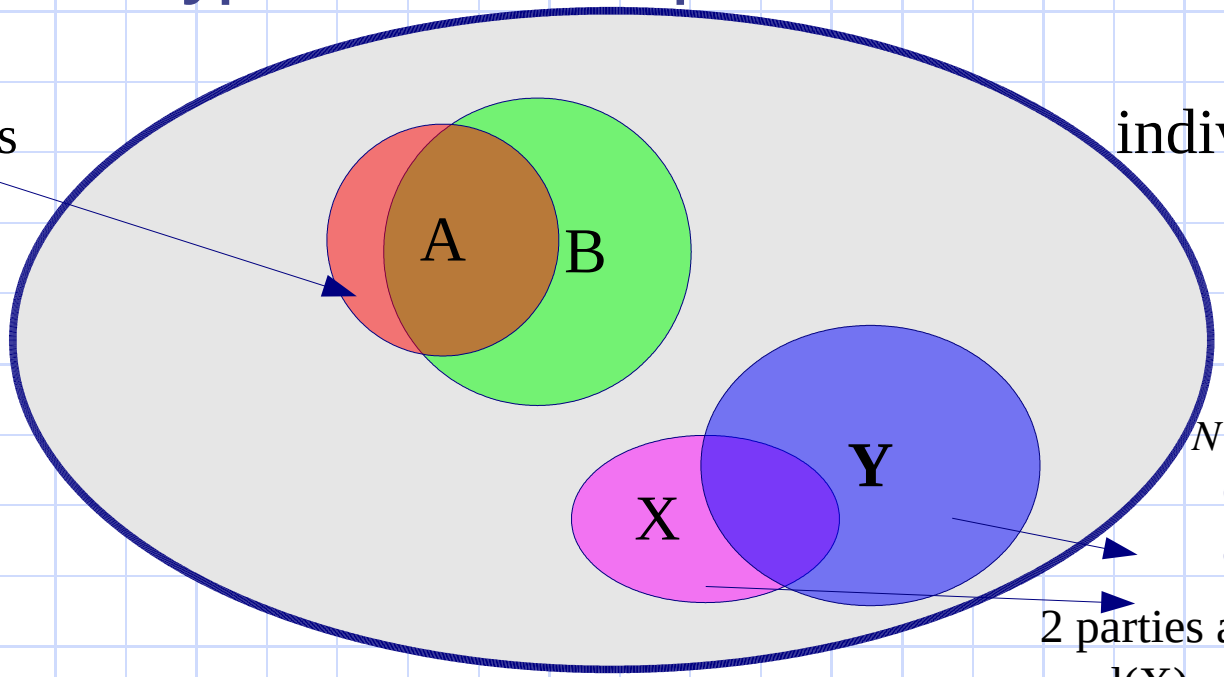
Modélisation des données

- On représente la base de textes comme ceci :



Rappel sur Intensité Implication

- Objectif : mesurer la qualité d'une règle $a \rightarrow b$
- Moyen : évaluer la rareté des contre-exemples dans les données sous hypothèse d'indépendance.



$$\varphi(a \rightarrow b) = 1 - Pr[N_{a\bar{b}} \leq n_{a\bar{b}}]$$

Évaluation d'implications

- On évalue toutes les implications du type :
  - terme → caractère
  - « Pour tout document, si le terme t est présent alors le document a tendance à être décrit par le caractère c»

	caractères
termes	φ(terme → caractère)

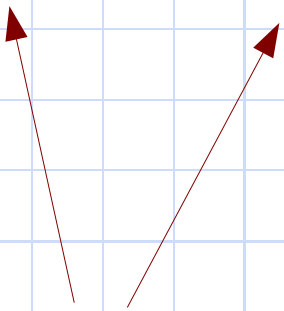
- indices :
  - support
  - intensité d'implication

$$\varphi(t \rightarrow c) = 1 - Pr[N_{t\bar{c}} \leq n_{t\bar{c}}]$$

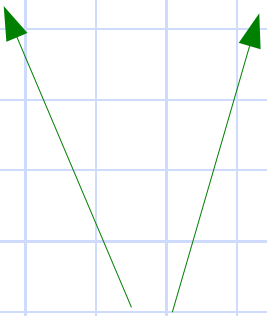
Association termes/caractères

Exemple de table d'implication :

φ(terme → caractère)	Extraversion	Extraversion moyenne	Rigueur	Dynamisme intellectuel
conscience professionnelle	0.0	0.63	0.99	0.0
sens de la méthode	0.77	0.0	0.92	0.0
preuve de créativité	0.0	0.0	0.0	0.94
attrait de la nouveauté	0.0	0.0	0.0	0.94
domaine de la communicatior	0.0	0.0	0.86	0.86



caractères ayant
de « mauvais » résultats



caractères ayant
de « bon » résultats

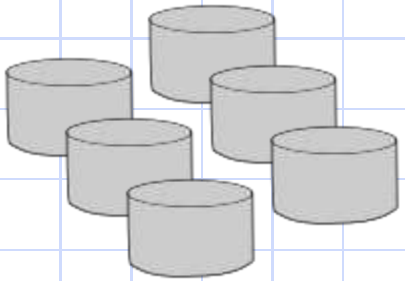
Association termes/caractères

- Formation des groupes -

- Associer pour chaque caractère, un ensemble de termes descriptifs.
  - un terme t est associé au caractère « E+ » si $\varphi(t \rightarrow E+) > seuil$
    - Pb : le choix du seuil
    - on choisit par défaut 0,5 (indépendance)

termes	caractères
	ϕ(terme,caractère)

seuillage



g groupes des termes
(1 groupe / caractère)

Résultats

- On obtient 30 ensembles de termes (1 ensemble par caractère psychologique)
  - exemple : CON+ (Rigueur)
    - seuil=0,5
    - nombre de termes sélect. = 25
    - moyenne des intensités = 0,93

Terme	φ(terme → CON+)
conscience professionnel	1,00
sens rigueur	0,99
personne confiance	0,98
qualité travail	0,97
preuve précision	0,97
réserve domaine	0,96
preuve rigueur	0,96
moyen oeuvre	0,94
fer gant	0,94
effort inutile	0,94
souci sécurité	0,94
fond chose	0,93
responsabilité gestion	0,93
sens méthode	0,92
élément régulation	0,90
ligne projet	0,90
méthode susceptible	0,90
manière précis	0,90
organisation temps	0,90
angle nouveau	0,90
manière réfléchir	0,90
solution nouveau	0,89
manière essentiel	0,89
atout important	0,86
domaine communication	0,86

Validation / jugement des résultats

- Méthode :
  - Pour chaque caractère C, on a un ensemble de termes G.
  - L'expert classe les termes en 2 groupes G_1 et G_2
    - G1 : les termes qui sont en adéquation avec le caractère
    - G2 : les termes qui ne sont pas en adéquation avec le caractère
  - A partir de cette classification, on en déduit la précision :

$$Précision(G) = \frac{card(G_1)}{card(G)}$$

Validation / jugement des résultats

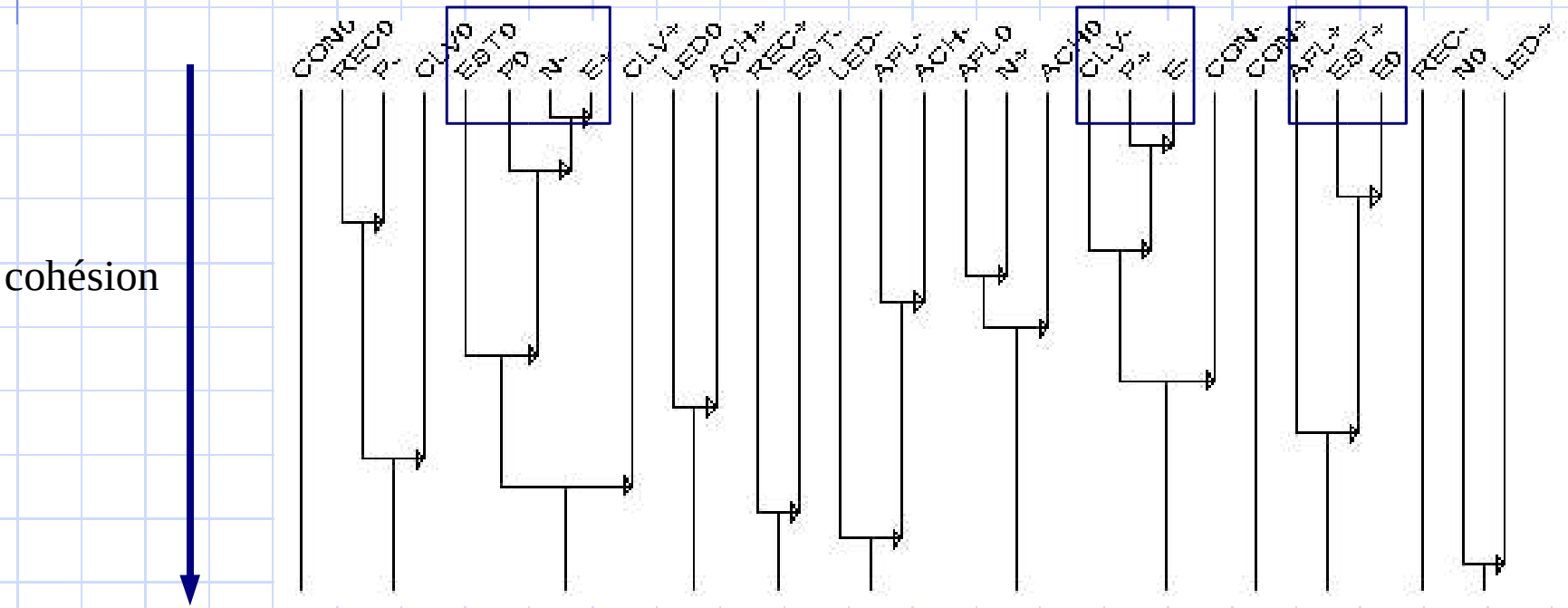
Caractère	Précision
Rigueur	1
Combativité	0.9
Anxiété	0.9
Dynamisme intellectuel	0.9
Affirmation	0.9
Remise en cause	0.9
Motivation de pouvoir	0.9
Motivation de protection	0.9
Détente	0.8
Improvisation	0.8
Motivation d'appartenance	0.8
Conciliation	0.7
Motivation d'indépendance	0.7
Anxiété moyenne	0.6
Conformisme intellectuel	0.6
Introversion	0.5
Extraversion	0.4
Extraversion moyenne	0.0

Groupes de termes représentatifs du caractère étudié.

Caractères non exprimés directement dans le texte mais servant à nuancer d'autres caractères

Etude des regroupements

- Un terme peut appartenir à plusieurs groupes
  - essayer d'évaluer les intersections entre les groupes.
  - mesurer les implications entre groupes de termes



Conclusion

- Méthode permettant d'associer des termes à des descripteurs d'un système d'indexation/catégorisation.
- Permet à l'expert d'étudier la modulation de son discours en fonction des individus étudiés.
- Bonne précision sur de nombreux caractères.

Perspectives

- Utiliser une approche de type k-means pour la formation des groupes de termes.
- Utilisation pour mesurer le degré d'appartenance d'instances à un concept.
  - élaborer une mesure de cohésion de concept ?
  - mesurer des distances entre concepts ?

FIN

Merci de votre attention

Bonne Journée

Problèmes

- Réglages des paramètres.
  - sélection des termes
  - seuil d'intensité pour former les groupes
- Explication des hiérarchies cohésitives.