Vers une identité européenne ?
Analyse des discours politiques en France, Grande-Bretagne et Allemagne dans les programmes électoraux
(1979-2004).

La lexicométrie nous permet-elle de comparer des corpus multilingues ?

Ronny Scholz, doctorant en co-tutelle
Institute de Sociologie – Université de Magdebourg – Allemagne
Université Paris-Est – CEDITEC

Démarche adoptée

- Questions et Hypothèses
- Présentation des corpus des programmes électoraux
- Comment comparer les corpus multilingues ?
  - Les problèmes de comparaison
  - Les propositions de solutions pour une comparaison des corpus multilingues
- L'exemple des Spécificités
- L'exemple des Fréquences relatives
- Conclusions

Questions de recherche

- Le projet de recherche analyse le développement d‘une identité européenne dans les programmes électoraux des élections européennes allemandes, françaises et britanniques (1979 – 2004).
- Quelles propositions pour une identité européenne peut-on trouver dans les discours politiques nationaux ?
- Quelles différences entre les des différents pays ? Quels points communs ?

Hypothèses

- Hypothèse: *Europe* existe seulement avec une signification variable dans les différents discours nationaux.

- Europe est un signifiant flottant dans les différents discours politiques. C'est-à-dire *Europe* reste toujours un signifiant arbitraire qui reçoit une signifiance avec chaque énonciation dans un certain contexte discursif (Laclau/Mouffe 1985).

Partis politiques dont les programmes électoraux forment le corpus étudié pour la période 1979 – 2004

Corpus allemand:

REP*, CDU, CSU, FDP, Die Grünen, SPD, PDS*

Corpus français:

FN, MPF, RPF, UMP (RPR, DL), UDF*, Les Verts*, PS, PRG*, PCF, LO*

Corpus britannique:

UKIP*, CONS, LDP, LAB, GREENS, PC*, SNP*

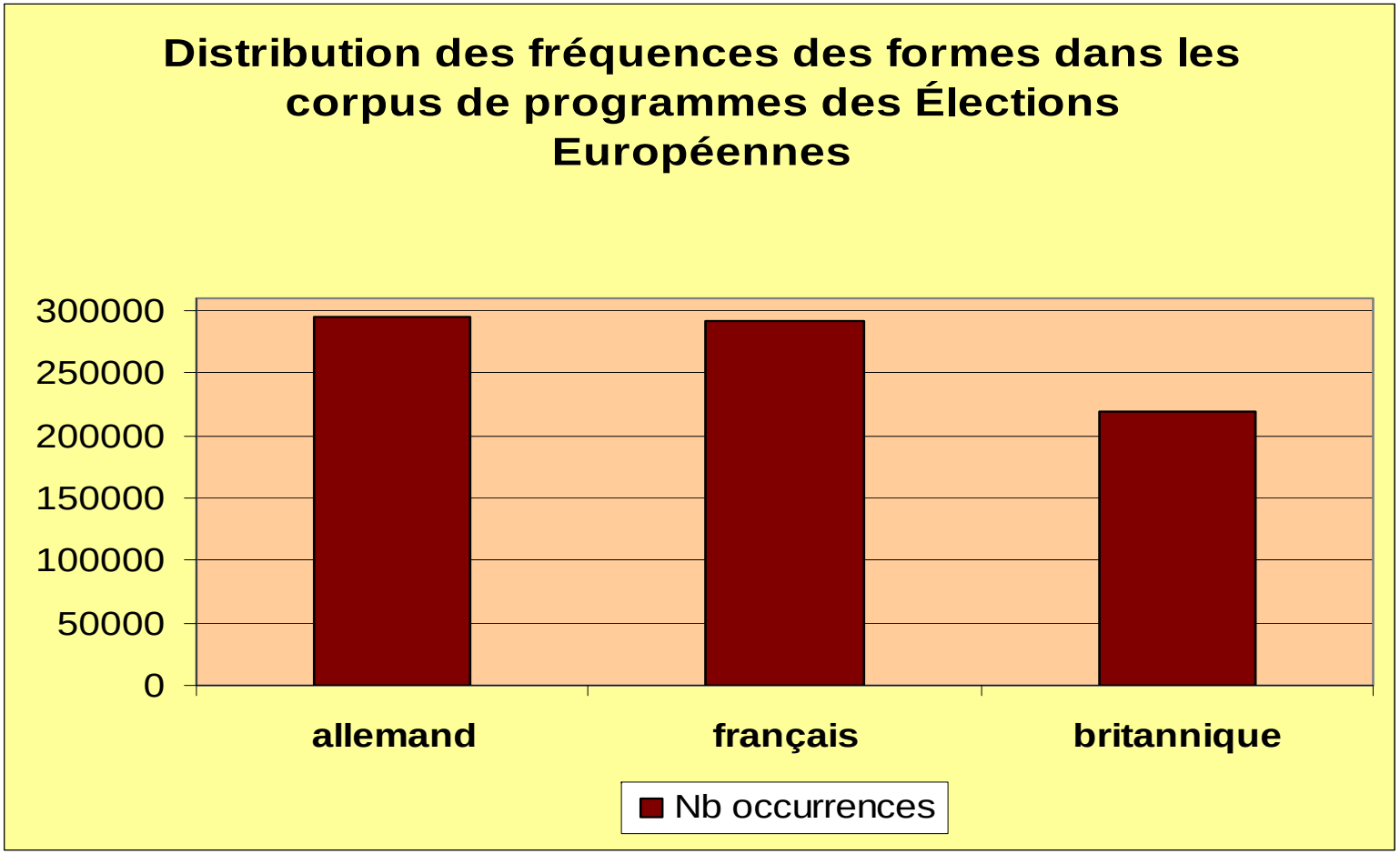
* Pour certains campagnes électorales aucun texte pouvait être étudiés du fait d'archives incomplètes ou inexistence du parti au moment d'élection étudié.

Caractéristiques des corpus

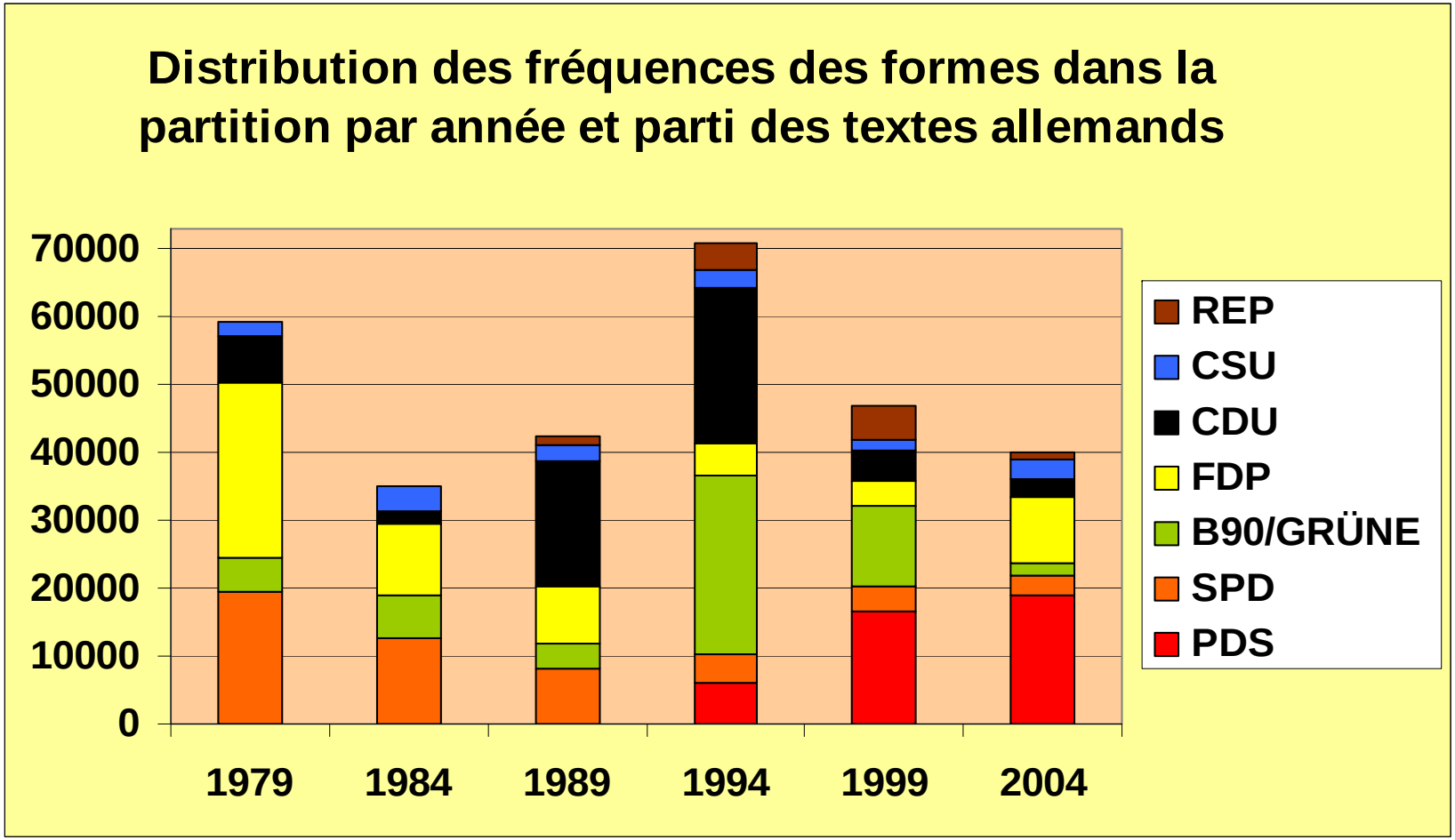
Corpus	Nb occurrences	Nb formes	Nb hapax
allemand	**294644**	24193	12644
français	**291497**	15703	6675
britannique	**219126**	9735	3663

Corpus	Fréq. Max	Forme	Nb Textes
allemand	15384	die	**37**
français	16373	de	**48**
britannique	15049	the	**36**

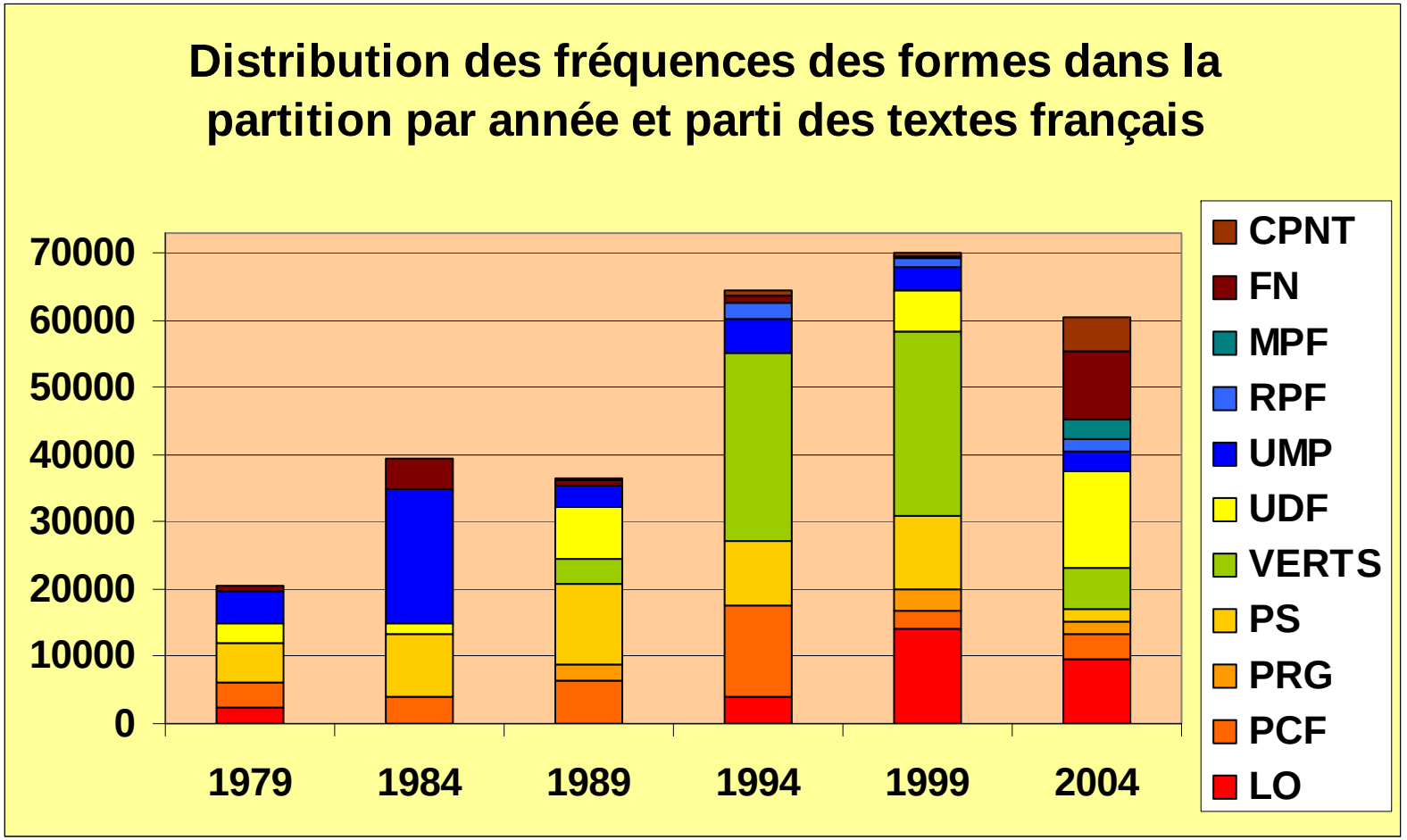
Présentation des corpus



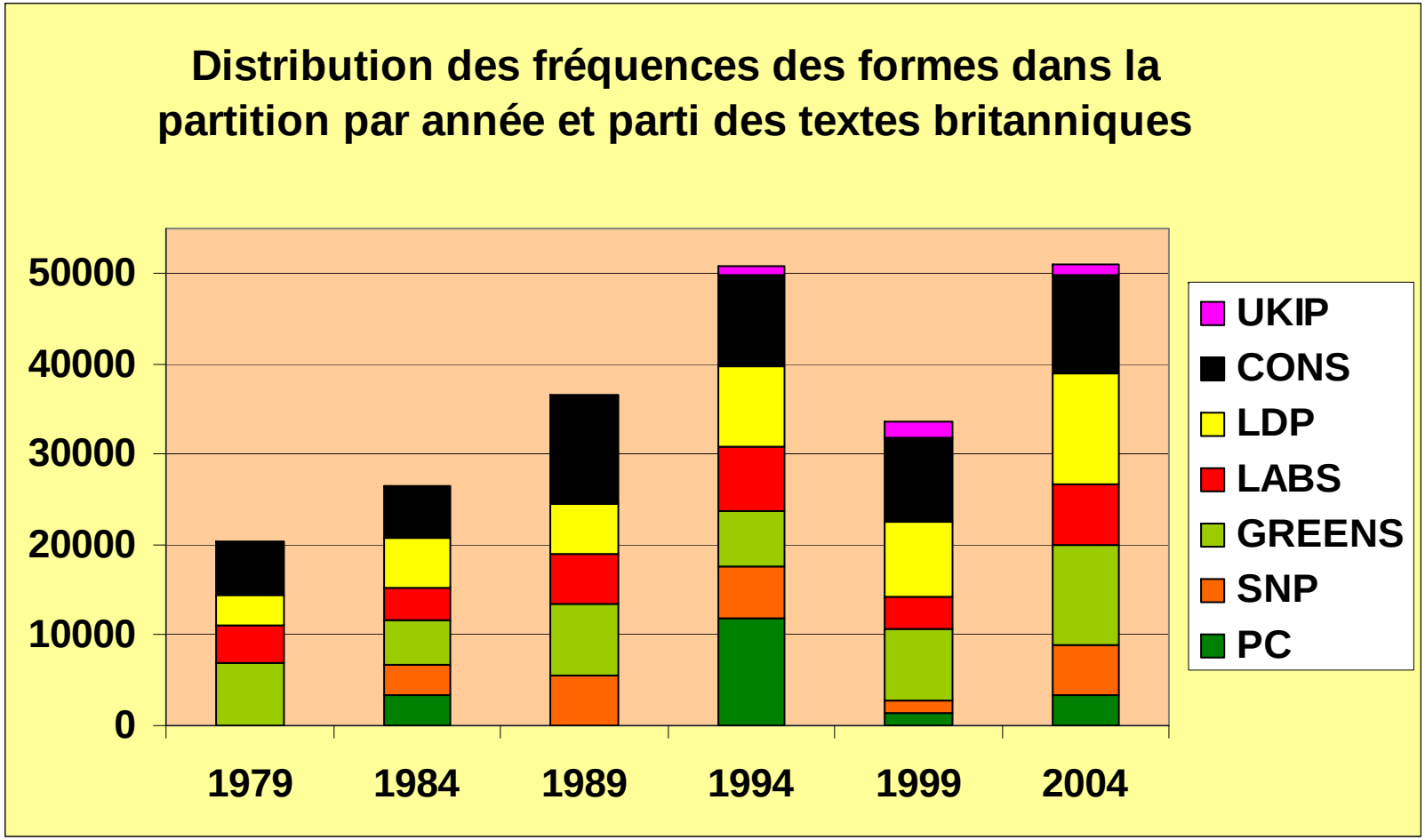
Présentation du corpus allemand
Nombre de textes : 37



Présentation du corpus français
Nombre de textes : 48



Présentation du corpus britannique
Nombre de textes : 36



Comment analyser l'identité européenne avec la lexicométrie

Les mesures lexicométriques comme:

- l'*index*,
- les *spécificités*,
- le *vocabulaire commun*,
- les *cooccurrences,*
- les *segments répétés*

révèlent des listes de vocabulaire qui permettent de tirer des conclusion sur le cotexte lexical dans lequel la notion d'*Europe* est développée.

Comparaison lexicométrique des corpus multilingues – PROBLÈMES au niveau des langues

Les propriétés sémantique (richesse vocabulaire) et morphologique d'une langue influencent systématiquement la statistique lexicométrique.

Cette influence pose des problèmes pour la comparaison lexicométrique des corpus multilingues. Il y aura notamment des différences à l'échelle de l'index des formes même dans un corpus traduit.

<u>Au niveau sémantique:</u>

- Problème de traduction des différents notions
  - Par exemple: *administration* → *Verwaltung*, *Behörde* (autorité)

<u>Au niveau morphologique</u> (problème de structure langagière différente):

- Les cas en allemand:

Europa / *Europas* **vs.** *Europe*

- Mots composés:

Osteuropa, *Westeuropa* **vs.** *Europe de l'Est, Europe de Ouest*

- La négation:

nicht **vs.** *ne pas*

- Flexions

speak **vs.** *parle, parles, parlons …* **vs.** *spreche, sprichst ...*

Comparaison lexicométrique des corpus multilingues – SOLUTIONS au niveau de la langue

- En comparant les trois corpus il faut toujours prendre en compte les particularités langagières qui influencent la statistique. Une différence entre des corpus peut toujours être un effet des différences de langues.
- MAIS: La tâche de la lexicométrie est d'expliquer les différences dans la langage des différents locuteurs. Dans un corpus monolingue on explique ces différences par les conditions institutionnelles et discursives différentes.
- Dans la recherche présentée ici le contexte institutionnel des trois corpus est très homogène. Tous les textes sont publiés par des partis politiques nationaux qui sont candidats pour le même parlement, le Parlement Européen.
- La différence de la langue est juste une dimension supplémentaire pour expliquer une différence dans la langage politique de plusieurs locuteurs.

Comparaison lexicométrique des corpus multilingues – PROBLÈMES au niveau de l'INDEX

La comparaisons des différents corpus sur la base d'index est douteux si :

- Dans un corpus constitué de textes venant de **contextes institutionnels variés**, l'index risque d'être surchargé du vocabulaire des textes d'une institution particulière.

  - En revanche, un corpus de textes venant d'un **seul contexte institutionnel** peut constituer une collection de textes représentatifs de cette institution et permettre de tirer des conclusions sur les particularités langagières de cette institution ou d'un discours particulier. Dans le cas présenté ici c'est le langage du discours politique.

La comparaisons des différents corpus sur la base d'index est douteux si :

- on compare les formes à **basses fréquences** des différent corpus – Ces formes ne représentent pas forcement le corpus entier et peuvent être introduit par les locuteurs secondaires.
- le corpus est **trop petit** - les anomalies lexicales d'un corpus ne sont pas équilibrées par la haute fréquence des formes qui sont utilisées fréquemment dans tous les textes d'un corpus.

- PROPOSITION: Un corpus est équilibré si la fréquence relative du nom le plus fréquent dépasse un certain nombre – peut être 20 (?) – dans la majorité des textes constituant le corpus. Plusieurs corpus équilibrés sont comparables au niveau de l'index des formes à hautes fréquences.

Les Fréquences et Spécificités de la forme Europa / Europe dans la partition <année>

	F relativ	F absolut	Spécificités	Longuer du parti
1984				
français	65	**260**	**-4**	40000
britannique	80	215	0	26000
allemand	66	235	+5	35000
1989				
français	95	350	+5	35000
britannique	82	300	0	35000
allemand	64	**275**	**+5**	42000
1994				
français	55	**350**	**-18**	65000
britannique	110	560	+18	50000
allemand	50	**350**	**0**	70000
1999				
français	80	**560**	**0**	**70000**
britannique	87	300	+2	34000
allemand	45	215	-3	46000
2004				
français	95	**580**	**+8**	**60000**
britannique	65	310	-7	40000
allemand	41	210	0	50000

Comparaison lexicométrique des corpus multilingues – PROBLÈMES au niveau des SPÉCIFICITÉS

- Chaque corpus a ses propriétés spécifiques (longueur des parties, fréquence d'une forme dans une partie)
- La mesure des spécificités dans les différentes partitions d'un corpus est une mesure relative entre les textes d'un corpus ou les différentes parties de ce corpus.
- Cette mesure représente les particularités du rapport entre toutes les parties d'un corpus.
- Comme les spécificités sont une mesure basée sur les particularités d'un corpus, elles ne sont pas comparables à l'échelle statistique avec d'autres corpus.

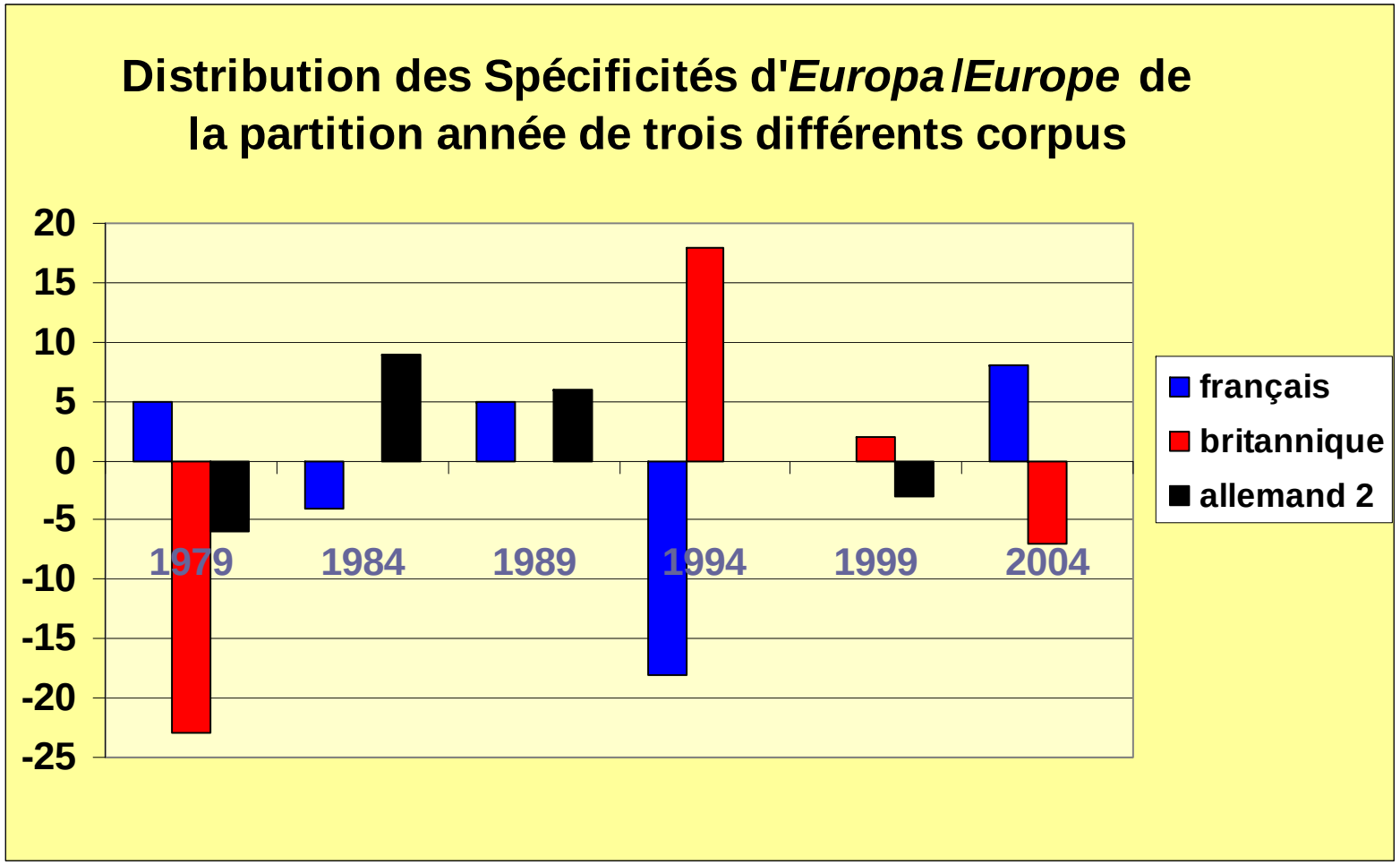
Comparaisons lexicométrique des corpus multilingues – SOLUTIONS au niveau des SPÉCIFICITÉS

- Les spécificités dans les différents corpus représentent des **tendances lexicales de ces corpus**. Dans une certaine mesure, ces tendances lexicales à l'intérieur du corpus peuvent être interprétés en tant que **réactions discursives dans un certain contexte socio-historique**.

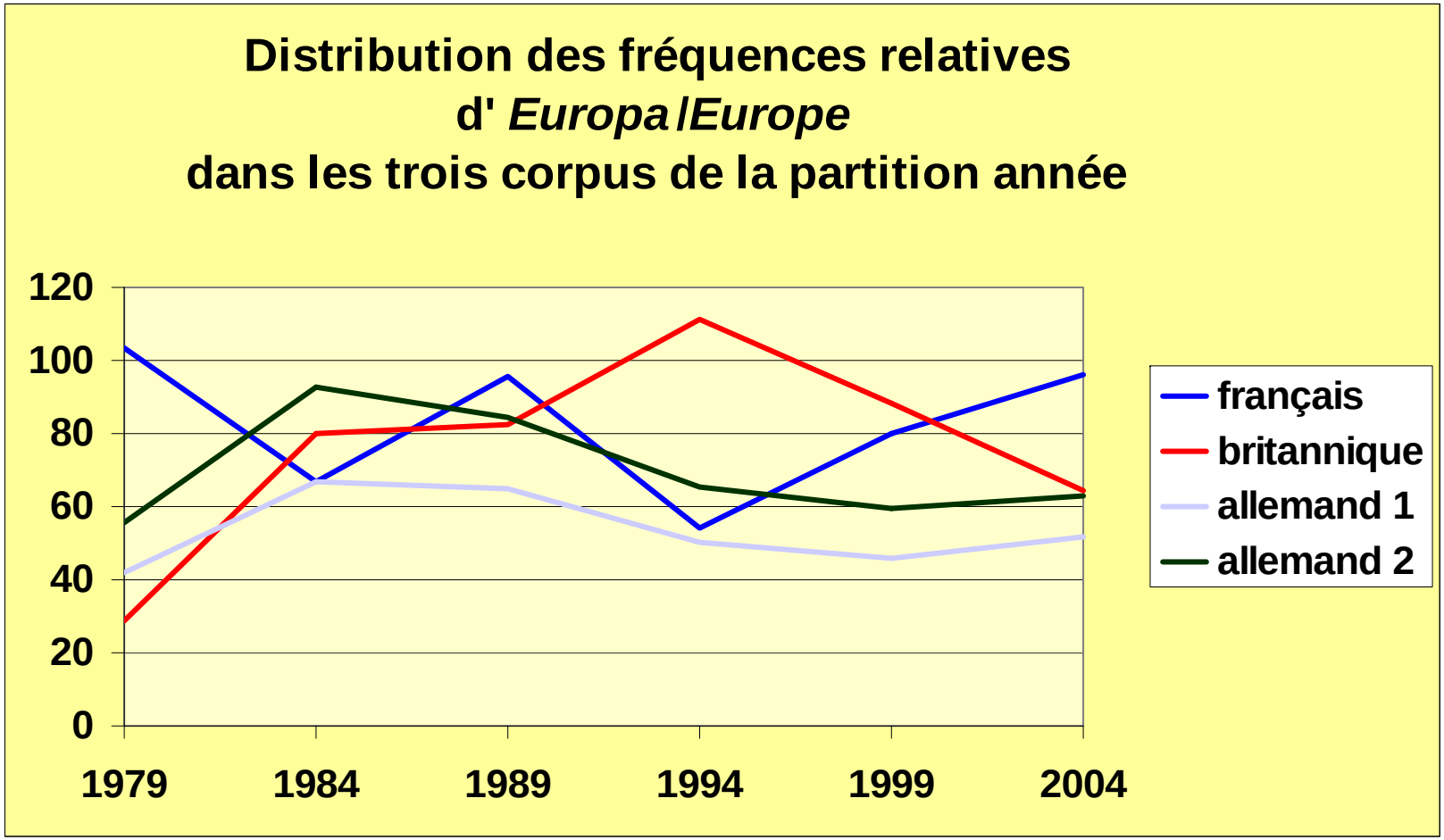
- Dans la dimension **diachronique** notamment, il y a une interaction entre le contexte socio-historique et la production des textes politiques.
  - Par exemple on peux se demander dans quelle mesure l'événement « Première élection directe du Parlement Européen » influence le contenu des programmes électoraux de ces mêmes élections ? Ou quelle influence a la ratification du traité d'Union Européen en 1993 sur le lexique des programmes électoraux de l'année suivante ? Comment la notion d'Europe est utilisée en 1994 dans les trois différents corpus? (*voit diapo suivant*)

- La comparaison des spécificités entre plusieurs corpus dans la dimension **synchronique** est un peu plus douteux mais aussi possible. On peut par exemple comparer les spécificités des différents partis politiques en tant que **comparaison des différents espaces discursifs**.
  - Dans ce sens on pourrait par exemple conclure que dans l'espace discursif allemand, le parti communiste a tendance à sur-employer la notion d'Union Européenne et de sous-employer la notion d'Europe. Tandis que dans l'espace discursif français les partis communistes sous-emploient les deux notions.

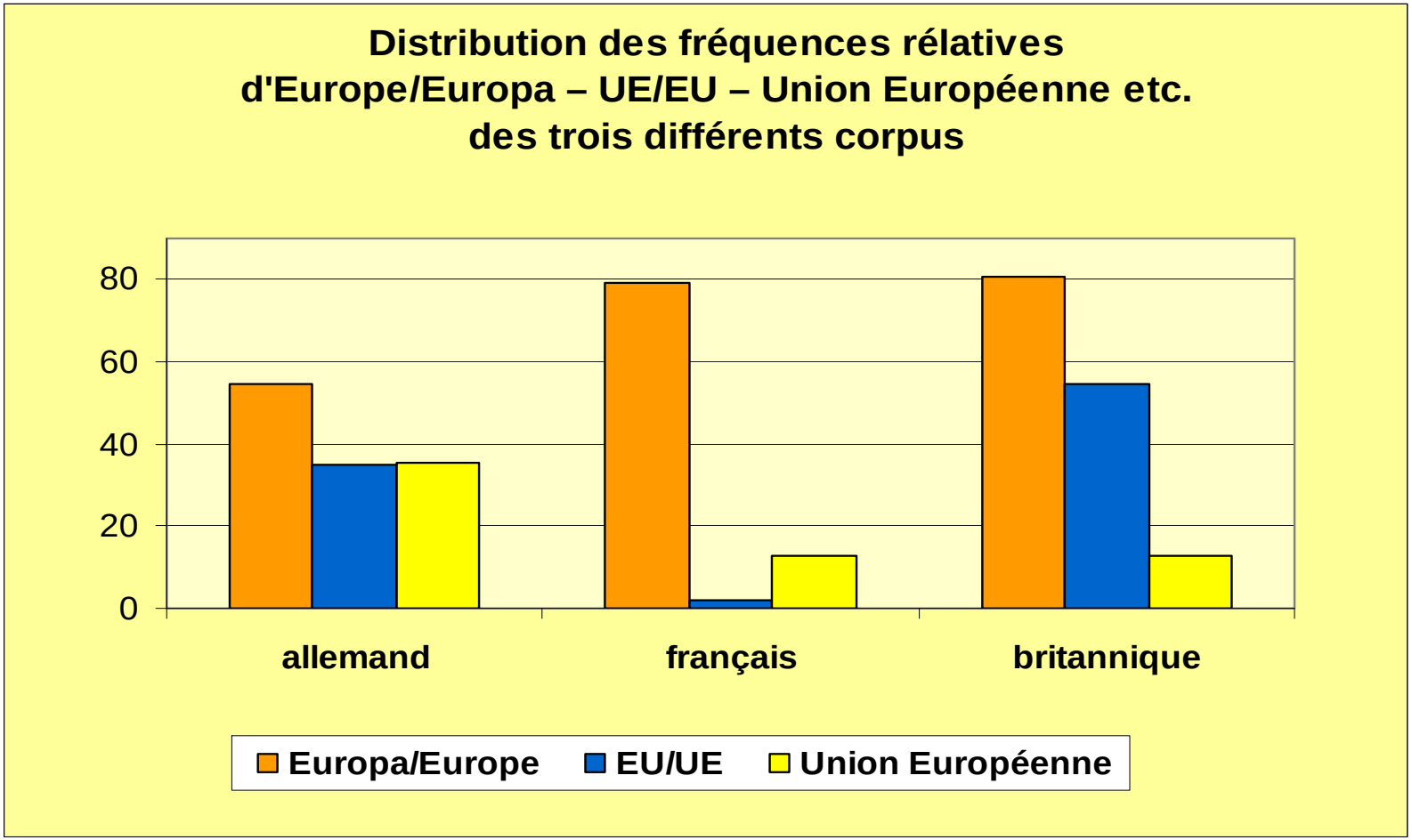
Tendances d'emploi spécifique dans la dimension diachronique de la notion d'Europe comparées dans les trois corpus



Les fréquences relatives – une comparaison statistique directe entre plusieurs corpus



Les fréquences relatives – une comparaison statistique directe entre plusieurs corpus



Conclusions – Index

- Dans une certaine mesure, l'analyse lexicométrique permet la comparaison **des corpus multilingues.**
- Une comparaison des ***fréquences relatives*** est toujours statistiquement correcte. Dans chaque cas particulier, il doit être décidé si cette comparaison est raisonnable sur le plan scientifique.
- Par ailleurs les corpus multilingues sont comparables :
  - A condition que les textes constituant le corpus soient produits dans un **contexte institutionnelle comparable**.
  - A partir de l'***index*** à condition que l'on compare les formes **à haute fréquence** dans un « corpus équilibré ».
  - Au niveau des ***spécificités*** si l'on focalise la comparaison sur les **tendances intracorporelles**. Par contre, une comparaison directe des valeurs des spécificités entre les corpus n'a pas de sens.

Merci beaucoup

pour votre

Attention!

Vielen Dank

für Ihre

Aufmerksamkeit!

Thank you very much

for your

attention!