Battre son corpus tant qu'il **ait** chaud

Benoît Habert
LIR – LIMSI – CNRS & université Paris X – Nanterre
habert@limsi.fr
http://www.limsi.fr/Individu/habert/

Plan

- Dégrouper automatiquement les sens
- Mesurer les changements distributionnels
- Battre son corpus : varier distance, contexte, partition
- Précautions et perspectives

Dégrouper les sens

- Réduire/repérer polysémie et homonymie
  - une source de bruit en recherche d'information (obtention de documents non pertinents)
  - un obstacle en acquisition sémantique (« ponts » indus entre regroupements)
  - l'indice de désaccords/méconnaissances dans les débats citoyens
- Difficulté : repérer les « mots qui en cachent d'autres »
- Dimensions
  - **repérer**/caractériser
  - **corpus partitionné**/non partitionné

Corpus et contextes

- Opposition
  - signalétique externe : LMP – 3 rubriques dans *Le Monde Parole*
  - interprétation : TCDT
    - 3 émetteurs CFTCa(vant), CFTCm(aintenue) et CFDT
    - tournants : scission entre CFDT et CFTCm ; pour la CFDT mai 68 puis 1979
- Contextes
  - syntaxiques : Syntex [Bourigault & Fabre 00]
  - « graphiques » : noms, verbes, adjectifs, adverbes dans une fenêtre
    - $\pm$ 5 mots
    - phrase

LMP ART ECO POL

Rubrique	*articles*	*mots*	*moyenne*	*maximum*
ART(s et médias)	2 261	1 080 620	477	2 990
ECO(nomie)	3 478	1 443 923	415	3 058
POL(itique)	2 305	1 326 576	575	5 202
Total	8 044	3 851 119		

TCDT AFC

Délimitation	*nom*	*mots*
CFTC jusqu'à la scission de 1964 incluse et CFDT d'avant mai 1968	TC45-64_DT65-67	29 704
CFTC maintenue	TC65-90	31 673
CFDT « radicalisée »	DT70-76	25 596
CFDT « recentrée »	DT79-92	34 677
Total (CFTC CFDT 1945-1992)	TCDT	121 650

Relations produites par Syntex 1/2

TCDT Phrase de la résolution générale CFDT 1965

Depuis la mise en place, en 1950, de ce ***salaire minimum interprofessionnel garanti en dessous*** duquel aucun **travailleur** de plus de 18 ans ne doit **être payé**, la **protection sociale** qui **présidait à** cette **institution s'est** sérieusement **dégradée**.

Relations produites par Syntex 2/2

Lemme 1	*POS 1*	*relation*	*lemme 2*	*POS 2*
salaire	N	EPI	minimum	A
salaire	N	EPI	interprofessionnel	A
salaire	N	EPI	garantir	A
garantir	V	en	dessous	N
payer	V	OBJ	travailleur	N
se dégrader	V	SUJ	protection	N
se dégrader	V	SUJ	protection social	S
présider	V	SUJ	protection	N
présider	V	SUJ	protection social	S
protection	N	EPI	social	A
présider	V	à	institution	N

Proximités distributionnelles

V dans *OBJ__travailleur*	partie	#	N dans *OBJ_payer*	partie	#
concerner	DT70-76	1	travailleur	DT70-76	1
défendre	DT70-76	1	travailleur	TC45-64_DT65-67	2
garantir	TC45-64_DT65-67	1	action	TC45-64_DT65-67	1
maintenir	DT70-76	1			
payer	DT70-76	1			
payer	TC45-64_DT65-67	2			
rendre	TC45-64_DT65-67	1			

« Colorer » les occurrences selon la partie

• Une occurrence dans une partie est suffixée du nom de cette partie

travailleur dans TCDT AFC $\longrightarrow$ *travailleur__TC45-64_DT65-67*, *travailleur__DT70-76*, *travailleur__DT79-92* et *travailleur__TC65-90*

• On peut mélanger sans les confondre ces hétérographes artificiels et leurs contextes

Exemple

Hétérographe artificiel	*contexte*	*fréquence*
travailleur__TC45-64_DT65-67	OBJ__payer	1
payer__TC45-64_DT65-67	OBJ__travailleur	1

Visualiser/mesurer les changements de distribution

- Visualisation : projection des hétérographes dans l'espace des contextes
  - proximité *financier* (≈ ? convergence)
  - dispersion *économique* (≈ ? divergence)
  - « bande à part » *travailleur, réel*
- Distance cumulée entre hétérographes d'un même lemme

2 distances

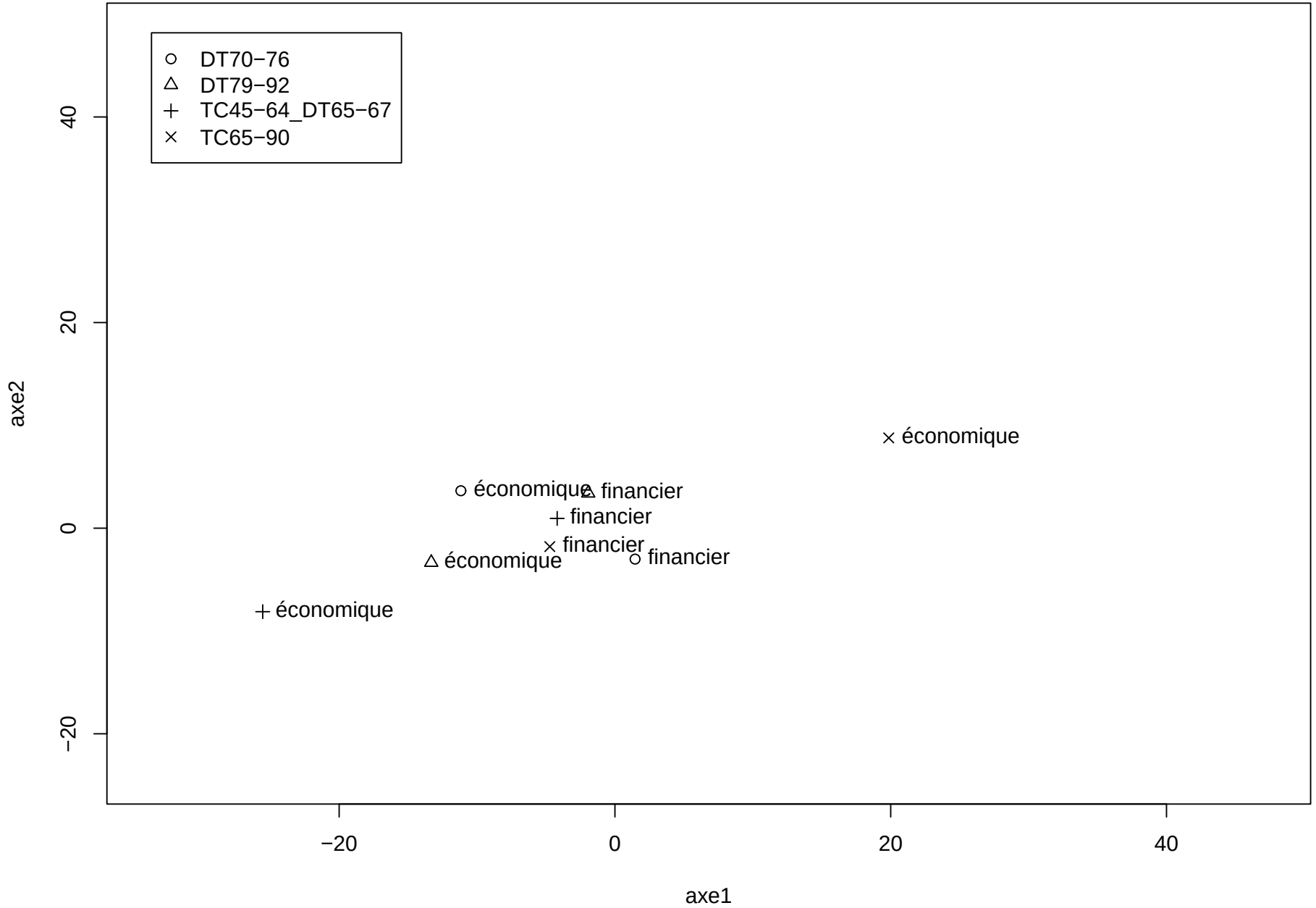
  - cosinus + pondération des traits (tf.idf)
  - Jaccard (sans pondération)

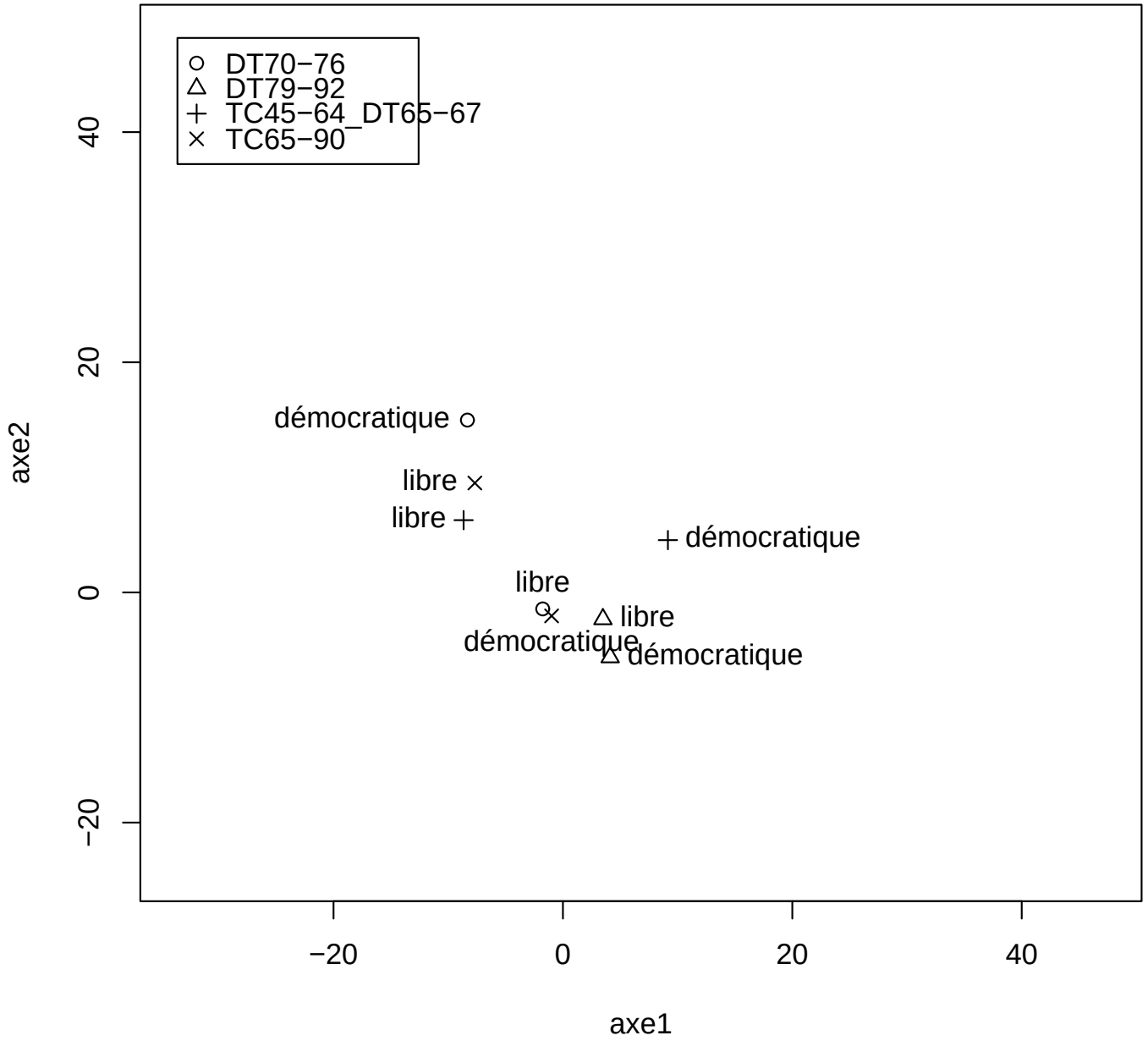
TCDT distances (cosinus, Jaccard) des hétérographes de *travailleur*

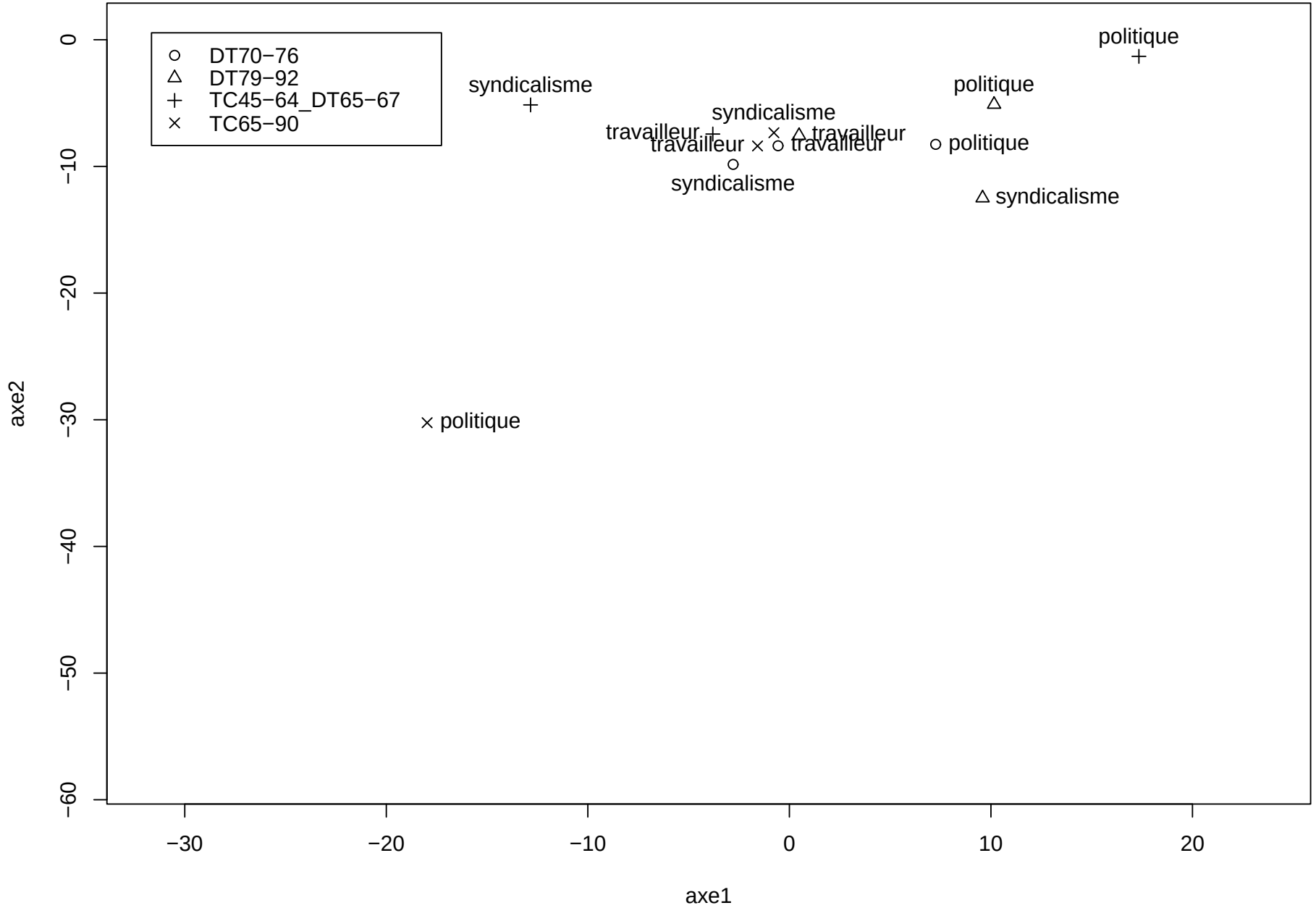
Hétérographe 1	*hétérographe 2*	*cosinus*	*Jaccard*
~__DT70-76	~__DT79-92	0.93	0.949
~__DT70-76	~__TC45-64_DT65-67	0.902	0.958
~__DT70-76	~__TC65-90	0.936	0.935
~__DT79-92	~__TC45-64_DT65-67	0.972	0.986
~__DT79-92	~__TC65-90	0.973	0.957
~__TC45-64_DT65-67	~__TC65-90	0.83	0.93
Somme des distances pour *travailleur*		5.543	5.715

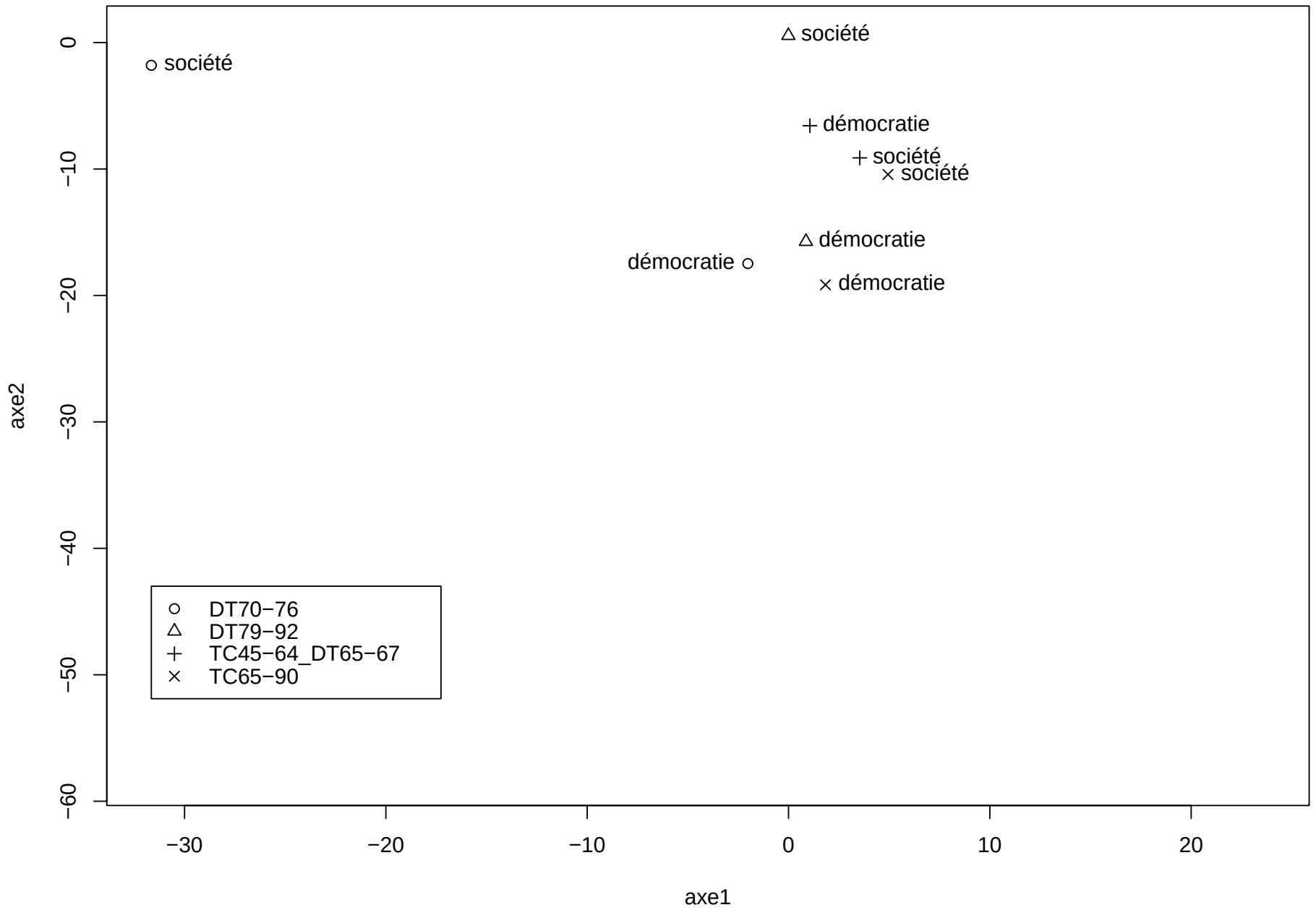
TCDT 10 premiers N partagés par distance cumulée entre hétérographes

Lemme	*cosinus*	*Jaccard*
ensemble	5.659	5.695
syndicalisme	5.623	5.734
travailleur	5.543	5.715
institution	5.515	5.68
relation	5.391	5.311
intérêt	5.34	5.564
solidarité	5.335	5.327
condition	5.329	5.432
situation	5.326	5.516
niveau	5.256	5.696


DT70−76
DT79−92
TC45−64_DT65−67
TC65−90
économique
financier
financier
financier
financier
économique
économique
économique
axe1
axe2
−20
0
20
40
−20
0
20
40


DT70−76
DT79−92
TC45−64_DT65−67
TC65−90
démocratique
libre
libre
démocratique
libre
libre
démocratique
démocratique
axe2
axe1
−20
0
20
40
−20
0
20
40


DT70−76
DT79−92
TC45−64_DT65−67
TC65−90
politique
syndicalisme
politique
syndicalisme
travailleur
travailleur
travailleur
politique
syndicalisme
syndicalisme
politique
axe2
0
−10
−20
−30
−40
−50
−60
−30
−20
−10
0
10
20
axe1


société
société
démocratie
société
société
démocratie
démocratie
démocratie
DT70−76
DT79−92
TC45−64_DT65−67
TC65−90
axe2
axe1
0
−10
−20
−30
−40
−50
−60
−30
−20
−10
0
10
20

Résultats LMP (Syntex + cosinus)

POS	*lemmes*
A	même, célèbre, classique, indispensable, fameux, possible, issu, traditionnel, présent, directeur, japonais, italien, militaire, complet, indépendant, urbain, allemand, populaire, permanent, rare
N	résultat, retour, titre, film, voix, salle, construction, auteur, évolution, autorité, cours, ouverture, gestion, demande, image, mouvement, maison, dirigeant, vente, ligne
V	remplacer, accompagner, découvrir, regrouper, naître, demeurer, apparaître, autoriser, continuer, se retrouver, entrainer, sembler, risquer, envisager, lier, construire, considérer, travailler, conduire, diriger

Résultats TCDT AFC (Syntex + cosinus)

POS	*lemmes*
A	véritable, nécessaire, financier, actuel, européen, essentiel, international, libre, politique, confédéral, mondial, tout, national, démocratique, nationale, économique, réel, local, professionnel, syndicaux
N	ensemble, syndicalisme, travailleur, institution, relation, intérêt, solidarité, condition, situation, niveau, moyen, société, effort, démocratie, entreprise, responsabilité, objectif, activité, emploi, politique
V	assurer, renforcer, devoir, confirmer, contribuer, développer, ouvrir, améliorer, assumer, réduire, apporter, poser, répondre, poursuivre, participer, réaffirmer, affirmer, continuer, engager, réaliser

Variation distributionnelle $\not\approx$ divergence

- Variation des contextes de verbes et d'adjectifs, voire de noms, ne correspondant pas à des différences d'acceptions
- Adjectifs
  - « de rangement » *social*, *économique*, *européen*, *mondial*
  - de nationalité *japonais*, *italien*, *allemand*
  - évaluatifs *indispensable*, *fameux*, *célèbre*, *véritable*
- Verbes
  - copules *sembler*, *apparaître*, *demeurer*
  - de parole *mandater*, *réaffirmer*, *rappeler*
  - dénotant le changement *poursuivre*, *améliorer*, *développer*
- Noms « outils » *niveau*, *ensemble*, *situation*, *moyen*

Variations dans LMP 1/2

- Artefact : distance cumulée forte par emploi écrasant dans une partie *film* (ART : 1 200 contextes / 25 POL / 21 ECO)
  - ART *auteur, salle*
  - ECO *cours*
  - POL *mouvement*
- Emploi prépondérant (mais déséquilibré) dans 2 parties
  - ECO > POL *autorité*, *construction*, *demande*, *évolution*, *gestion*, *résultat*
  - POL > ECO *dirigeant*
  - ART > POL *image, voix*
  - ART > ECO *maison*
  - ECO > ART *ouverture, titre, vente*

Variations dans LMP 2/2

- Spécialisations
  - *cours* monnaie, Bourse (ECO) / processus (ART)
  - *mouvement* collectif (POL) / déplacement (ART) et évolution (ART, ECO)
  - *image* concret / abstrait (ART donne plus de place au premier sens que POL et ECO)
  - *ouverture* début d'un morceau de musique (ART)
  - *voix* qualité, tessiture (ART) / votes (POL)

Variations dans TCDT AFC 1/3

- Unité en thème et en genre du corpus $\longrightarrow$ distributions moins iniques que LMP
- Associations dominantes
  - DT79-92 et TC65-90 *activité*, *emploi*, *entreprise*
  - TC45-64_DT65-67 et DT70-76 *démocratie*
  - DT70-76 et DT79-92 *société*
- *condition*
  - contextes peu révélateurs *EPI__fondamental*, *OBJ__améliorer*, *EPI__favorable*, *EPI__indispensable*, *EPI__essentiel*, *EPI__nécessaire*
  - opposition entre DT70-76 - DT79-92 OBJ (c'est quelque chose sur laquelle agir) et TC65-90 - TC45-64_DT65-67 avant tout EPI ou SUJ

Variations dans TCDT AFC 2/3

- *démocratie*

TC45-64_DT65-67 / DT70-76 renvoi à une menace *OBJ_restaurer, OBJ_rétablir*

- *société*
  - partages à 4 *EPI__démocratique* ou 3 *OBJ__construire*
  - TC45-64_DT65-67 et TC65-90 *pour__homme*
  - DT70-76 et DT79-92 *EPI__autogestionnaire*, *EPI__capitaliste*, *EPI__inégalitaire*, *EPI__socialiste* et *OBJ__changer*
  - DT70-76 contextes très politiques *EPI__aliénant*, *EPI__autogérer*, *EPI__autogéré*, *EPI__décentraliser*, *EPI__futur*, *EPI__hiérarchique*, *EPI__technocratique*, *à base de__autogestion*
  - DT79-92 *EPI__ouvert*, *EPI__pluraliste*, *EPI__pluriculturel*, *EPI__répressif*, *EPI__éclaté*

Variations dans TCDT AFC 3/3

- *syndicalisme*
  - sauf TC65-90 *EPI__démocratique*
  - TC65-90 et TC45-64_DT65-67 *EPI__chrétien*, *EPI__libre*
  - DT70-76 peu de contextes propres
  - TC65-90 - TC45-64_DT65-67 SUJ privilégié par rapport à DT79-92.
  - échos *EPI__constructif* TC45-64_DT65-67 / *EPI__responsable* TC65-90

- *travailleur*
  - *SUJ_participer* 3 parties sur 4
  - TC45-64_DT65-67 et TC65-90 *EPI_chrétien* et *EPI_manuel*
  - TC65-90 faible emploi net
  - DT79-92 fragmentation de la catégorie *EPI_français*, *EPI_immigré*, *EPI_originaire*, *EPI_syndiqué*, *EPI_permanent*, *EPI_engagé*
  - DT70-76 représentation plus conflictuelle *OBJ_défendre*, *en_lutte*, *en_situation_de_subordination*, *face_à_pouvoir*.
  - DT79-92 et DT70-76 favorisent SUJ sur OBJ.

Varier les distances – LMP

Distance	
cosinus	résultat, retour, titre, **film**, voix, salle, construction, auteur, évolution, autorité, cours, ouverture, gestion, demande, image, mouvement, maison, dirigeant, vente, ligne
Jaccard	musique, **film**, théâtre, voix, musée, hausse, droite, femme, reprise, compagnie, oeuvre, achat, chaine, scène, histoire, cours, art, actionnaire, opposition, maire

Varier les distances – TCDT AFC

Distance	
cosinus	**ensemble**, **syndicalisme**, **travailleur**, **institution**, relation, **intérêt**, solidarité, condition, situation, **niveau**, **moyen**, **société**, **effort**, **démocratie**, **entreprise**, **responsabilité**, **objectif**, activité, **emploi**, politique
Jaccard	**syndicalisme**, **travailleur**, **emploi**, **niveau**, place, **ensemble**, **institution**, service, **objectif**, **moyen**, **société**, pouvoir, **responsabilité**, **entreprise**, **effort**, **intérêt**, mise, **démocratie**, droit, prise

Varier les contextes

- Version du corpus étiquetée par/pour Syntex
- Restriction aux noms, adjectifs, verbes et adverbes (degré, nég.)
- Fenêtre graphique (ie sans relations syntaxiques)
- 2 versions
  - $\pm$ 5 mots (quelle que soit leur catégorie)
  - phrase
- Exemple pour *travailleur* dans la phrase-exemple
  - Syntex {*payer*/OBJ}
  - $\pm$ 5 mots {**dessous**}
  - phrase {mise, en_place, 1950, salaire, minimum, interprofessionnel, garantir, **dessous**, an, *payer*, protection, sociale, présider, institution, se_dégrader, sérieusement}

Les matrices se remplument

	Syntex	± 5 N, A, V, Adv	Phrase N, A, V, Adv
% occupé	0.25	0.61	2.94
Formes	271	438	438
Hétérographes	1 084	1 748	1 748
Fmax	232	692	4 217
moyenne	12.58	35.47	226
1er quart.	3.0	7.0	42.0
2e quart.	7.0	18.0	113.0
3e quart.	15.0	41.0	260.25
Contextes	3 648	4 318	4 746
Fmax	405	1163	6 517
moyenne	3.73	14.36	83.33

TCDT 4 parties – Contextes partagés par toutes les parties pour*travailleur*

Syntex – 0	
± 5 N, A, V, Adv	
ne ; **pays** ; **pouvoir** ; **être**	4
Phrase N, A, V, Adv	
action ; activité ; apporter ; assurer ; autre ; besoin ; collectif ; condition ; congrès ; croissance ; droit ; développement ; emploi ; en particulier ; entreprise ; faire ; grand ; immigré ; intérêt ; **ne** ; niveau ; nouveau ; nécessaire ; nécessité ; organisation ; participer ; pas ; **pays** ; plus ; politique ; **pouvoir** ; public ; responsabilité ; rôle ; salaire ; salarié ; secteur ; situation ; social ; structure ; syndicalisme ; tout ; travail ; travailleur ; vie ; volonté ; économie ; économique ; **être**	49

TCDT 4 parties – Nombre de contextes propres pour *travailleur*

	Syntex	$\pm$ 5	Phrase
DT70-76	24	67	159
DT79-92	26	59	147
TC45-64_DT65-67	36	94	374
TC65-90	11	41	145

TCDT 4 parties – Syntex – Nombre de contextes partagés pour *travailleur*

DT70-76	DT79-92	TC45-64_DT65-67	TC65-90
24	*26*	*36*	*11*
2	2		
2		2	
1			1
	1	1	
	1		1
		3	3
	0	0	0
1		1	1
1	1		1
0	0	0	
0	0	0	0

TCDT 4 parties – ± 5 N, A, V, Adv – Partages *travailleur*

DT70-76	DT79-92	TC45-64_DT65-67	TC65-90
67	*59*	*94*	*41*
15	15		
7		7	
5			5
	8	8	
	2		2
		9	9
	4	4	4
3		3	3
3	3		3
7	7	7	
4	4	4	4

TCDT 4 parties – Phrase N, A, V, Adv – Partages *travailleur*

DT70-76	DT79-92	TC45-64_DT65-67	TC65-90
159	*147*	*374*	*145*
53	53		
54		54	
18			18
	56	56	
	12		12
		71	71
	27	27	27
27		27	27
9	9		9
43	43	43	
49	49	49	49

Comparaison des distances et par type de contexte

Contextes	*lemmes des* k *premiers N par distance cumulée (cosinus) entre hétérographes*
Syntex	`ensemble, syndicalisme, travailleur, institution`, relation, `intérêt`, solidarité, condition, situation, `niveau, moyen, société, effort, démocratie, entreprise, responsabilité, objectif`, activité, `emploi`, politique
$\pm$ 5	*fait*, *intervention*, *exigence*, *en_vue*, occasion, domaine, base, accroissement, propriété, *projet*, perspective, insertion, institution, victime, *investissement*, *réalisation*, renforcement, système, état, utilité
Phrase	**histoire**, **assistance**, occasion, **intégration**, **dépendance**, **appareil**, **aspect**, **pensée**, dégradation, sauvegarde, esprit, **liaison, jour**, utilité, conjoncture, libéralisme, intéressé, connaissance, **adhésion**, nature

Contextes	*lemmes des* k *premiers N par distance cumulée (Jaccard) entre hétérographes*
Syntex	`syndicalisme, travailleur, emploi, niveau,` place, `ensemble, institution,` service, `objectif, moyen, société,` pouvoir, `responsabilité, entreprise, effort, intérêt,` mise, `démocratie`, droit, prise
± 5	revenu, *intervention*, *exigence*, perspective, chômage, part, *fait*, *réalisation*, réduction, *projet*, création, enseignement, propriété, *investissement*, institution, accroissement, sécurité, *en_vue*, nécessité, régime
Phrase	insuffisance, **histoire**, égard, capital, règle, licenciement, esprit, poursuite, **appareil**, disposition, **liaison, jour, dépendance**, **pensée**, nationalisation, **aspect**, communauté, **intégration**, **adhésion**, **assistance**

Varier les partitions

- Contrastes d'émetteurs : CFTCa(vant), CFTCm(aintenue), CFDT
  - CFTCa / CFTCm / CFDT
  - CFTCa / CFTCm
  - CFTCa / CFDT
  - CFTCm / CFDT
- En aval d'une interprétation : 4 parties issues de l'interprétation d'une analyse factorielle des correspondances

Résultats pour d’autres partitions

Partition lemmes des k *premiers N par distance (cosinus) entre hétérographes* (a : avant scission, m : maintenue – contexte : ± 5 N, A, V, Adv)	
TCa DT	perspective, partie, libération, exploitation, contact, compensation, productivité, collectivité, patronat, instauration, valeur, discussion, projet, effet, demande, profession, nature, adhésion, enfant, an
TCa TCm DT	partie, direction, nature, libération, combat, discussion, rémunération, projet, intervention, réalité, instauration, système, régression, statut, difficulté, autonomie, valeur, revalorisation, soutien, en_vue
TCa TCm	statut, généralisation, forme, an, partie, système, accroissement, réglementation, baisse, souci, direction, commission, syndicat, intervention, information, en_vue, nationalisation, contexte, méthode, devoir
TCm DT	obligation, lien, dialogue, coordination, présence, modalité, discussion, conception, source, partie, environnement, risque, égard, nationalisation, direction, argent, nature, concertation, thème, manière

Partition lemmes des k *premiers N par distance (Jaccard) entre hétérographes* (a : avant scission, m : maintenue – contexte : ± 5 N, A, V, Adv)	
TCa DT	pratique, perspective, adhérent, projet, patronat, capacité, part, temps, accord, autonomie, réalité, mobilisation, fonctionnement, exploitation, égalité, libération, profit, difficulté, collectivité, choix
TCa TCm DT	rémunération, partie, marché, direction, transformation, tendance, valeur, autonomie, accord, demande, temps, projet, répartition, intervention, grève, forme, combat, union, productivité, an
TCa TCm	accord, statut, syndicat, an, initiative, caractère, information, crise, rémunération, en_vue, marché, loi, temps, équipement, négociation, généralisation, forme, employeur, dépense, esprit
TCm DT	classe, conception, unité, proposition, fonctionnement, union, rémunération, mouvement, cotisation, coordination, discussion, application, fonds, débat, discrimination, conscience, adhérent, valeur, risque, reconnaissance

Caractériser les distributions

- Voisins dans la partie
  - changement des traits (ensemble des contextes de la partie pour la catégorie / ensemble des contextes partagés par les lemmes retenus)
  - les voisinages peuvent être divergents
- Examen exploratoire (AFC) des ensembles de contextes pour repérer des directions ($\not\approx$ acceptions)
- Examen exploratoire (AFC) du regroupement des ensembles de contextes selon les parties
- Partitionnement en k (arbitraire) classes des ensembles de contextes, puis examen (AFC) des regroupements
  - « individus »
    - les ensembles de contextes (ici la phrase)
    - les classes d'ensembles

Caractériser par les voisins : TCDT AFC *travailleur*

Partie	*Jaccard*
TC45-64_DT65-67	syndicat (0.039), organisation (0.031), **diplômé** (0.024), **étape** (0.024), **civilisation** (0.024), **ouvrier** (0.024), potentiel (0.023), **voie** (0.023), **morale** (0.023), concurrence (0.023).
DT70-76	ensemble (0.045), **prédétermination** (0.032), **prédominance** (0.032), **emprise** (0.032), **participe** (0.031), **cours** (0.031), adhésion (0.031), hégémonie (0.031), **militaire** (0.031), **cohérence** (0.031).
DT79-92	**économie** (0.062), mouvement (0.048), salarié (0.036), **imagination** (0.033), degré (0.032), **diversification** (0.032), refus (0.032), **masse** (0.032), particulier (0.031), réalisation (0.031).
TC65-90	jeu (0.083), syndicalisme (0.07), service (0.066), travail (0.057), entreprise (0.057), **prévention** (0.056), **isme** (0.056), **père** (0.056), établissement (0.053), **morale** (0.053).

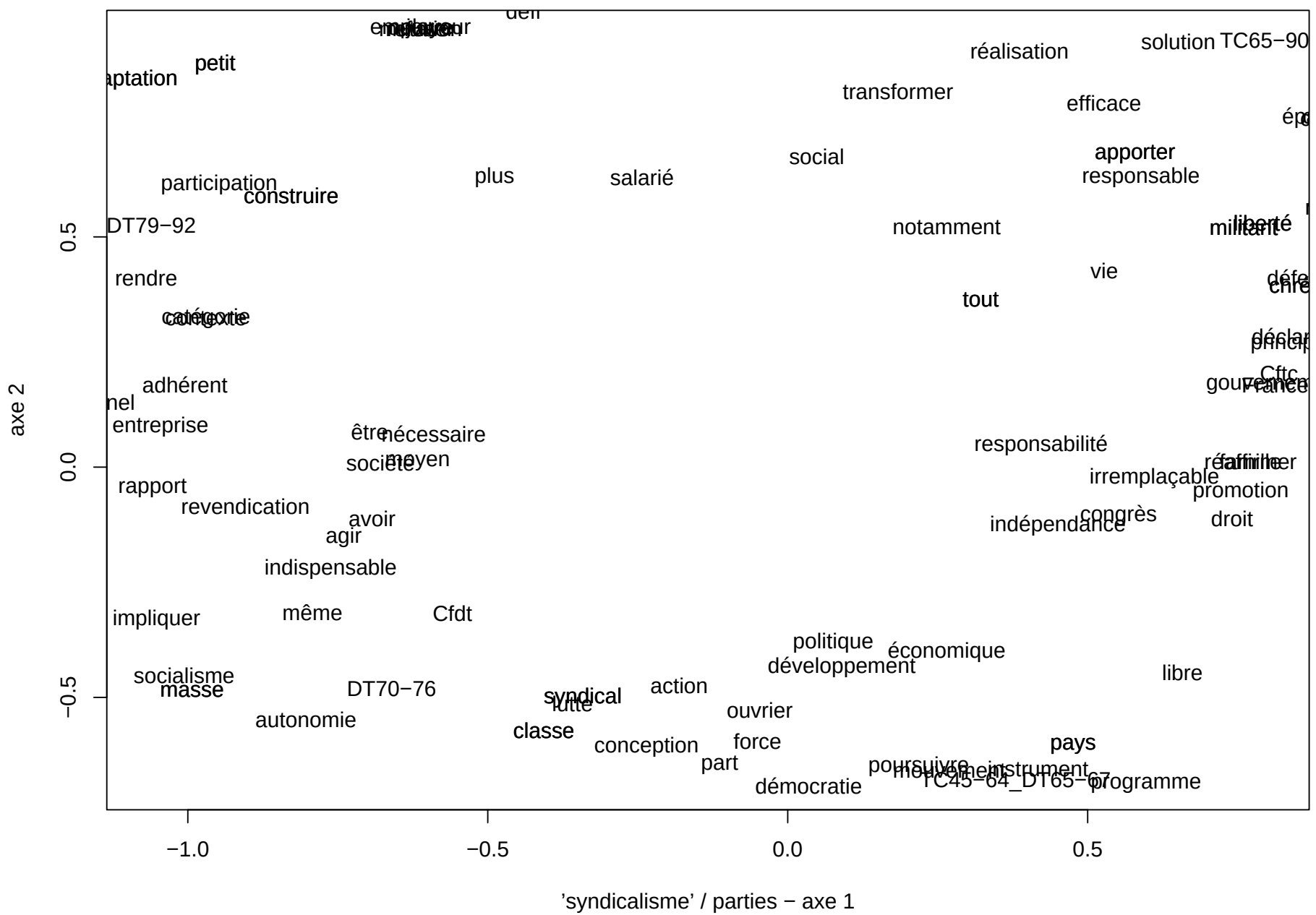
Voisins et distances : TCDT AFC *travailleur*

Partie	*Cosinus*	*Jaccard*
TC45-64_DT65-67	**civilisation** (0.187), **étape** (0.153), syndicalisme (0.139), personne (0.137), **morale** (0.133), **concurrence** (0.115), **diplômé** (0.113), **voie** (0.109), calcul (0.108), **ouvrier** (0.101).	syndicat (0.039), organisation (0.031), **diplômé** (0.024), **étape** (0.024), **civilisation** (0.024), **ouvrier** (0.024), potentiel (0.023), **voie** (0.023), **morale** (0.023), **concurrence** (0.023).

Explorations autour de *travailleur*

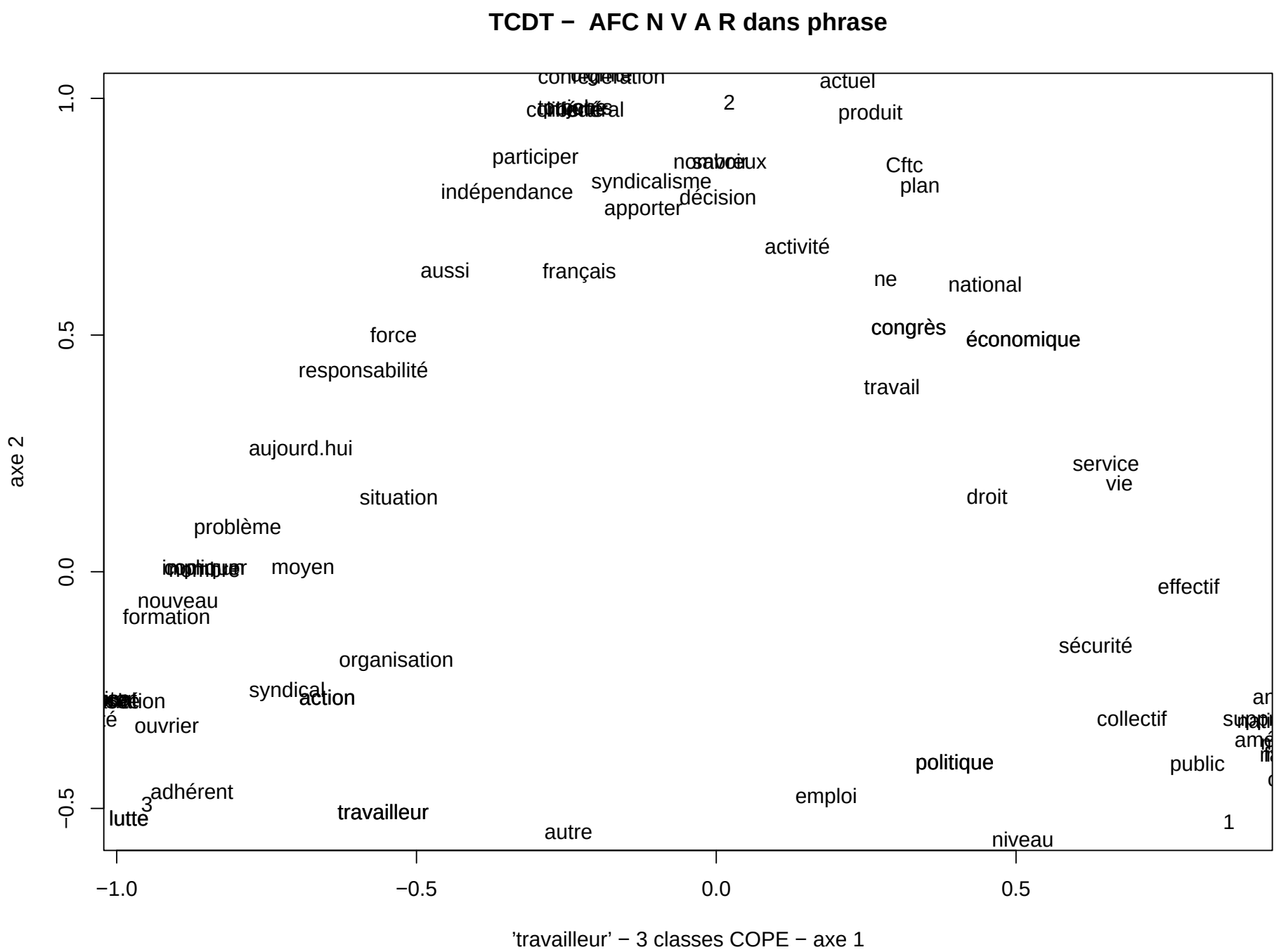
- AFC des 4 427 associations (1 408 contextes différents) de *travailleur* avec les noms, adjectifs, verbes et adverbes des 252 phrases où il figure (occupation : 1.19%, moy. phrase : 19.50, moy. contexte : 3.14)
- AFC des 4 parties (occupation : 38.37%, moy. partie : 1 106.75, moy. contexte : 3.14)
  - TC45-64_DT65-67 77 phrases, 1 929 associations (817 contextes)
  - DT70-76 67 phrases, 894 associations (462 contextes)
  - DT79-92 67 phrases, 959 associations (474 contextes)
  - TC65-90 41 phrases, 648 associations (408 contextes)
- AFC des *k* classes fournies par un classifieur

travailleur par parties


TCDT 4 parties − AFC N V A R dans phrase
petit
participation
construire
DT79−92
rendre
adhérent
entreprise
rapport
revendication
impliquer
socialisme
masse
autonomie
indispensable
avoir
agir
être
nécessaire
société
moyen
même
Cfdt
DT70−76
syndical
lutte
action
classe
conception
part
ouvrier
force
démocratie
politique
développement
économique
libre
pays
poursuivre
instrument
programme
TC45−64_DT65−67
responsabilité
irremplaçable
indépendance
congrès
promotion
droit
réalisation
solution TC65−90
transformer
efficace
apporter
responsable
social
plus
salarié
notamment
militant
liberté
vie
tout
gouvernement
France
Cftc
axe 2
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
'syndicalisme' / parties − axe 1

3 classes pour *travailleur*

contraste entre 3 classes pour *travailleur*


TCDT − AFC N V A R dans phrase
actuel
produit
participer
Cftc
plan
indépendance
syndicalisme
apporter
décision
activité
aussi
français
ne
national
congrès
économique
force
responsabilité
travail
aujourd.hui
service
vie
situation
droit
problème
moyen
effectif
nouveau
formation
sécurité
organisation
syndical
action
collectif
ouvrier
politique
public
adhérent
emploi
travailleur
lutte
autre
niveau
axe 2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
’travailleur’ − 3 classes COPE − axe 1

Se distancier des distances

- Pondérer en fonction de la taille du corpus et de sa redondance

(LMP/TCDT : faible taille + évolution idéologique sur 47 ans)

- Varier les distances pour contrebalancer la compulsion interprétative
- Distance cumulée des distributions $\not\equiv$ différence d'acception

Pallier l'émiettement des contextes

- Matrices « fantomatiques »
  - Fragmentation supplémentaire liée aux hétérographes artificiels
  - $\longrightarrow$ points de contact limités entre hétérographes artificiels
- Multiplier et « alourdir » les partages de contextes
  - Associer contextes syntaxiques/graphiques
  - Recherche ciblée de contextes (extraction d'information)
- Réduction de l'espace des traits
  - Calcul de similarités de second ordre
  - Projection de connaissances exogènes (synonymes, classes de mots)

Continuer à battre ses corpus

- Amélioration par rapport à une visualisation de la dispersion des hétérographes artificiels
- Objectif du repérage de « mots qui en cache d'autres » non atteint
  - instabilité des distances $\longrightarrow$ ordre fourni peu fiable
  - « brouillage » des décalages trop importants des répartitions dans les parties du corpus
  - plutôt inflexions de sens qu'écarts d'acceptions
- Ne pas attendre l'obésité des corpus : les corpus à « divergences » (forums, débats citoyens, controverses) ne sont pas forcément volumineux