

Le « poids » des langues

Aix-en-Provence 27-28 septembre
2007

Le « poids » des langues

Alain Calvet
Louis-Jean Calvet

Aix-en-Provence 27-28 septembre
2007

1532: Lettre de Gargantua à son fils

« J'entens et veulx que tu aprenes les langues parfaictement. Premierement la *grecque*, comme le veult Quintilian, secondement la *latine* ; et puis *l'hebraïcque* pour les saintes lettres, et la *chaldaïcque* et *arabicque* pareillement ; et que tu formes ton style quant a la grecque, a l'imitation de Platon ; quant a la latine, de Ciceron... »

2007: Enquête dans six universités de Rio de Janeiro

« Quelles langues aimeriez-vous qu'apprennent
vos enfants ? »

Anglais (82%)

Espagnol (30%)

Français (25%)

D'une façon générale

Comment mesurer l'importance relative des langues ?

Comment les classer ?

On pense en général au nombre de leurs locuteurs

Mais les différentes évaluations de ce nombre varient considérablement :

Prenons par exemple trois sources différentes

(chiffres de 2003)

	Quid	Linguasphere	S.I.L. (ethnologue)
1. Chinois	1 milliard	1 milliard	885 millions
2. Anglais	500 millions	1 milliard	322 millions
3. Hindi	497 millions	900 millions	182 millions
4. Espagnol	392 millions	450 millions	332 millions
5. Russe	277 millions	320 millions	170 millions
6. Arabe	246 millions	250 millions	---
7. Bengali	211 millions	250 millions	189 millions
8. Portugais	191 millions	200 millions	177 millions
9. Malais	159 millions	160 millions	23 millions
10. Français	129 millions	125 millions	72 millions
11. Allemand	128 millions	125 millions	98 millions
12. Japonais	126 millions	130 millions	125 millions

Aix-en-Provence 27-28 septembre
2007

Qui nous donnent des classements différents

Quid	Linguasphere	Ethnologue
Chinois	Chinois	chinois
Anglais	Anglais	Espagnol
Hindi	Hindi	Anglais
Espagnol	Espagnol	Bengali
Russe	Russe	Hindi
Arabe	Arabe	Russe
Bengali	Bengali	Portugais
Portugais	Portugais	Japonais
Malais	Malais	Allemand
Français	Japonais	Français
Allemand	Français	
Japonais	Allemand	

Aix-en-Provence 27-28 septembre
2007

Le « poids » des langues

Il s'agit d'effectuer une classification des langues du monde, fondée sur des facteurs discriminants, puis d'analyser plus finement cette classification en utilisant des méthodes statistiques d'analyse des données

Nous pouvons imaginer différents facteurs

Nombre de locuteurs « langue maternelle »

Nombre de locuteurs « langue seconde »

Traductions « source »

Traductions « cible »

Prix Nobel de littérature

Production cinématographique

Existe-t-il un traitement de texte?

Existe-t-il un correcteur orthographique?

Aix-en-Provence 27-28 septembre

2007

Ou encore ...

Peut-on consulter Google ou Yahoo dans cette langue?

Des articles dans Wikipedia sont-ils rédigés dans cette langue?

Nombre de pays dans lesquels la langue est officielle/co-officielle

Nombre de pays dans lesquels on peut étudier ces langues

Importance sur Internet

Indice de Développement Humain des pays dans lesquels
on parle ces langues

...Ou encore

Indice de fécondité des pays dans lesquels
on parle ces langues

Poids économique des pays dans lesquels
on parle ces langues

Importance de la langue dans les échanges économiques

Flux touristiques

Entropie

Etc ...

Un problème

Ces facteurs permettrait d'effectuer une classification des langues en deux groupes (valeurs catégorielles : oui/non) ou en une hiérarchie de 1 à 7000 (valeurs continues)

Mais tous ces facteurs ne nous donnent pas le même type d'informations. Le taux de développement humain par exemple varie de 0 à 1 tandis que le nombres de locuteurs varie de 1 à 800.000.000

Transformation linéaire des facteurs

Nous ramenons donc, pour chaque facteur, la valeur minimale à 0, la valeur maximale à 1 et appliquons une transformation linéaire pour les valeurs intermédiaires, ce qui permet d'affecter une importance "égale" à chacun des facteurs :

Transformation linéaire du nombre de locuteurs

Langue	Rang	Valeur absolue	Valeur normée
Mandarin	1	725.5	1
Portugais	2	174.5	0.240283
Haoussa	3	45	0.061729
Xiang	4	36	0.04932
Malayalam	5	35.8	0.049044
Bahasa	6	30.3	0.04146
Visayan/Cebuano	7	20	0.027259
Népalî	8	17.2	0.023398
Hongrois	9	13.1	0.017745
Slovaque	10	8	0.010713
Norvégien	11	4.5	0.005887
Islandais	12	0.23	0

Différentes classifications possibles

- Selon le rang pour chacun des facteurs pris individuellement (1, 2, 3, ...,N)
- *Selon la somme des rangs de chacun des facteurs*
- *Selon la somme des valeurs normées*
- *Selon une sélection plus limitée de facteurs*
- *Selon une combinaison pondérée ou non des classifications précédentes*
- *Selon des méthodes mathématiques plus évoluées*
- *Etc ...*

Dix facteurs ...

- Nombre de locuteurs
- Nombre de pays dans lesquels la langue a un statut officiel
- Nombre d'articles dans Wikipedia
- Nombre de prix Nobel de littérature
- Entropie
- Taux de fécondité
- Indice de développement humain (IDH)
- Taux de pénétration d'internet
- Nombre de traductions, langue cible
- Nombre de traductions, langue source

... et leurs sources

- Nombre de locuteurs et statut officiel :
 - <http://www.ethnologue.com/web.asp>
- Wikipedia
 - http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics
- Prix Nobel
 - <http://nobelprize.org/>
- Entropie
 - calculée à partir des données de population
- Taux de fécondité
 - <http://www.prb.org/FrenchContent.aspx>
- IDH
 - <http://www.undp.org/french/>
- Taux de pénétration d'internet
 - <http://www.internetworldstats.com/stats.htm>
- Index translationum :
 - <http://databases.unesco.org/xtrans/stat/xTransStat.html>

Entropie ?

- Entropie = $-\sum(p_i \cdot \text{Log}(p_i))$
- Langue parlée très majoritairement dans un seul pays
 - 0.98 $-p_i \text{Log} p_i = 0.0198$
 - 0.02 $-p_i \text{Log} p_i = 0.0782$ Entropie = 0.098
- Langue parlée dans plusieurs pays de démographies comparables
 - 0.34 $-p_i \text{Log} p_i = 0.367$
 - 0.33 $-p_i \text{Log} p_i = 0.366$
 - 0.33 $-p_i \text{Log} p_i = 0.366$ Entropie = 1.099
- Marathi : 0.003 Amharique: 0.019
- Espagnol : 2.509 Arabe : 2.279

Classification selon le nombre de locuteurs

- 1. Mandarin
- 2. Hindi
- 3. Anglais
- 4. Espagnol
- 5. Arabe
- 6. Portugais
- 7. Bengali
- 8. Russe
- 9. Japonais
- 10. Allemand
- 11. Pendjabi
- 12. Javanais
- 13. Wu
- 14. Vietnamien
- 15. Tagalog
- 16. Tamoul
- 17. Min
- 18. Coréen
- 19. Français
- 20. Marathi

Autre classification, selon le facteur « langue officielle »

- 1. Anglais
- 2. Français
- 3. Arabe
- 4. Espagnol
- 5. Portugais
- 6. Allemand
- 7. Italien
- 8. Russe
- 9. Bahasa
- 10. Néerlandais *
- 10. Hongrois*
- 10. Mandarin*
- 10. Roumain*
- 10. Farsi*
- 10. Croate*
- 10. Slovène*
- 10. Albanais*
- 10. Tamoul*
- 10. Swahili*
- 10. Bambara*

Aix-en-Provence 27-28 septembre
2007

*: A partir du néerlandais toutes
ces langues sont officielles dans
3 pays

Selon la présence des langues sur Internet

- 1. Anglais
- 2. Mandarin
- 3. Espagnol
- 4. Japonais
- 5. Allemand
- 6. Français
- 7. Coréen
- 8. Italien
- 9. Portugais
- 10. Malais
- 11. Néerlandais
- 12. Arabe
- 13. Polonais
- 14. Suédois
- 15. Thaï
- 16. Turc
- 17. Russe
- 18. Vietnamien
- 19. Farsi
- 20. Roumain

Selon le nombre de prix Nobel de littérature

- 1. Anglais 24
- 2. Français 12
- 3. Allemand 11
- 4. Espagnol 10
- 5. Russe 5
- 6. Italien 5
- 7. Suédois 4
- 8. Polonais 4
- 9. Grec 2
- 10. Japonais 2
- 11. Danois 2
- 12. Ex-aequo : 1
Mandarin, Finnois,
Tchèque, Islandais,
Arabe, Provençal,
Yiddish, Portugais,
Bengali, Hébreu,
Hongrois, Turc, Serbo-
croate
- 24. Les autres 0

Les traductions (langue source)

- 1. Anglais
- 2. Francais
- 3. Allemand
- 4. Russe
- 5. Italien
- 6. Espagnol
- 7. Suedois
- 8. Danois
- 9. Tcheque
- 10. Neerlandais
- 11. Polonais
- 12. Japonais
- 13. Hongrois
- 14. Arabe
- 15. Norvegien
- 16. Portugais
- 17. Hebreu
- 18. Mandarin
- 19. Finnois
- 20. Bahasa

Comment croiser ces données?

Nous constatons que certaines langues (anglais, mandarin, espagnol, français....) se retrouvent dans le groupe de tête de la plupart des classements et que d'autres n'apparaissent que dans un seul classement (Yiddish, Grec et prix nobel).

Différentes classifications possibles

- *Selon le rang pour chacun des facteurs pris individuellement (1, 2, 3, ...,N)*
- Selon la somme des rangs de chacun des facteurs
- Selon la somme des valeurs normées
- *Selon une sélection plus limitée de facteurs*
- *Selon une combinaison pondérée ou non des classifications précédentes*
- Selon des méthodes mathématiques plus évoluées, analyse en clusters, analyse en composantes principales ...
- Etc ...

Deux classements sur dix facteurs

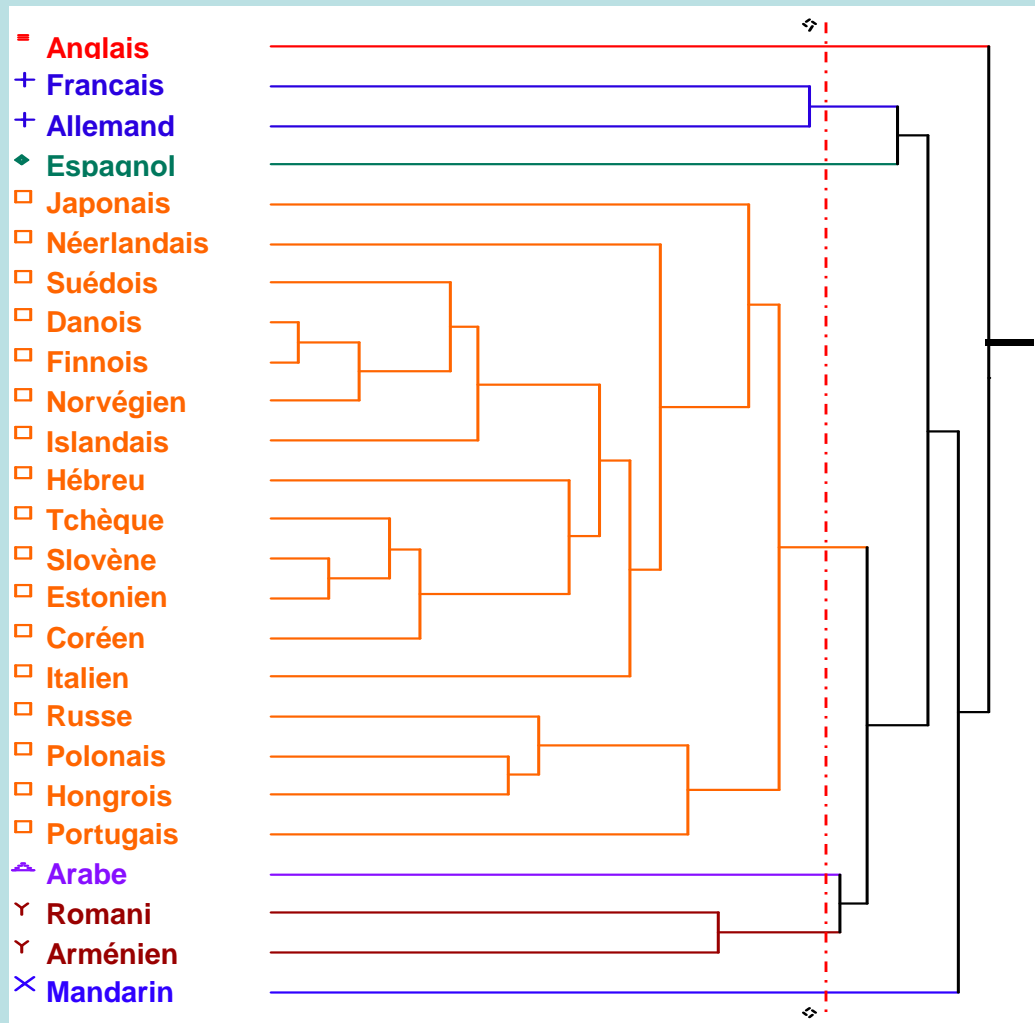
Somme des valeurs normées		Somme des rangs	
Anglais	Islandais	Anglais	Japonais (-6)
Français	Finnois	Français	Polonais (+ 3)
Espagnol	Romani	Espagnol	Danois (-3)
Allemand	Russe	Allemand	Hongrois (+8)
Japonais	Polonais	Néerlandais (+1)	Coréen (+8)
Néerlandais	Portugais	Russe (+8)	Turc (+10)
Arabe	Norvégien	Portugais (+9)	Finnois (-5)
Suédois	Mandarin	Italien (+1)	Tchèque (+2)
Italien	Hébreu	Arabe (-2)	Hébreu
Danois	Tchèque	Suédois (-2)	Serbe (+22)

Allons plus loin

A partir d'un nombre élevé de facteurs il est plus intéressant d'utiliser des méthodes spécifiques de traitement de données.

Par exemple, pour les 25 premières langues voici une analyse en clusters :

Analyse en clusters : un exemple



- Certains clusters sont des singletons, d'autres des doubletons et enfin un nombre élevé de langues se regroupent en un seul cluster
- Un singleton indique pour la langue concernée une singularité par rapport à un ou plusieurs facteurs
- Pour tester ce comportement nous avons créé de fausses langues en modifiant systématiquement les paramètres les uns après les autres

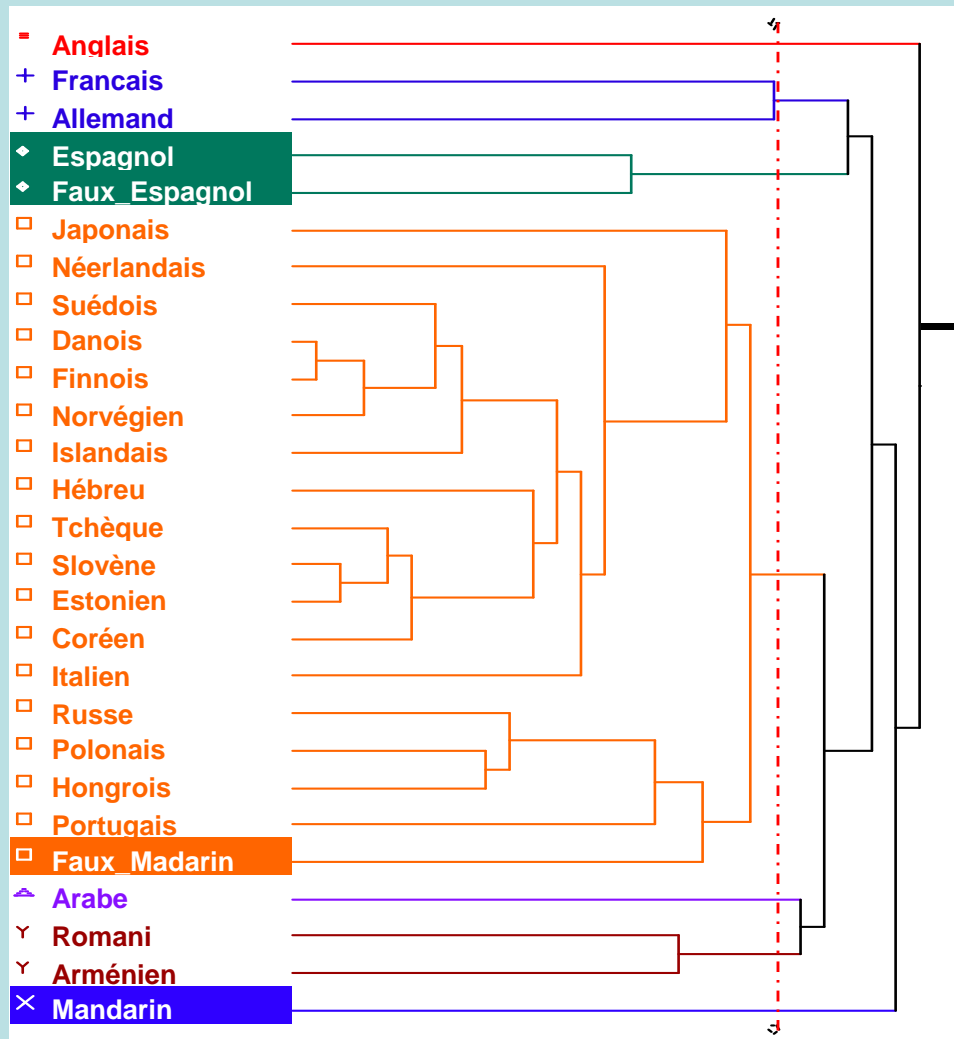
Deux fausses langues

Un « faux » espagnol a été créé en modifiant le nombre de pays où il est langue officielle : 1 pays au lieu de 21

Un « faux » mandarin en modifiant le nombre de locuteurs : 50 millions de locuteurs au lieu de 800

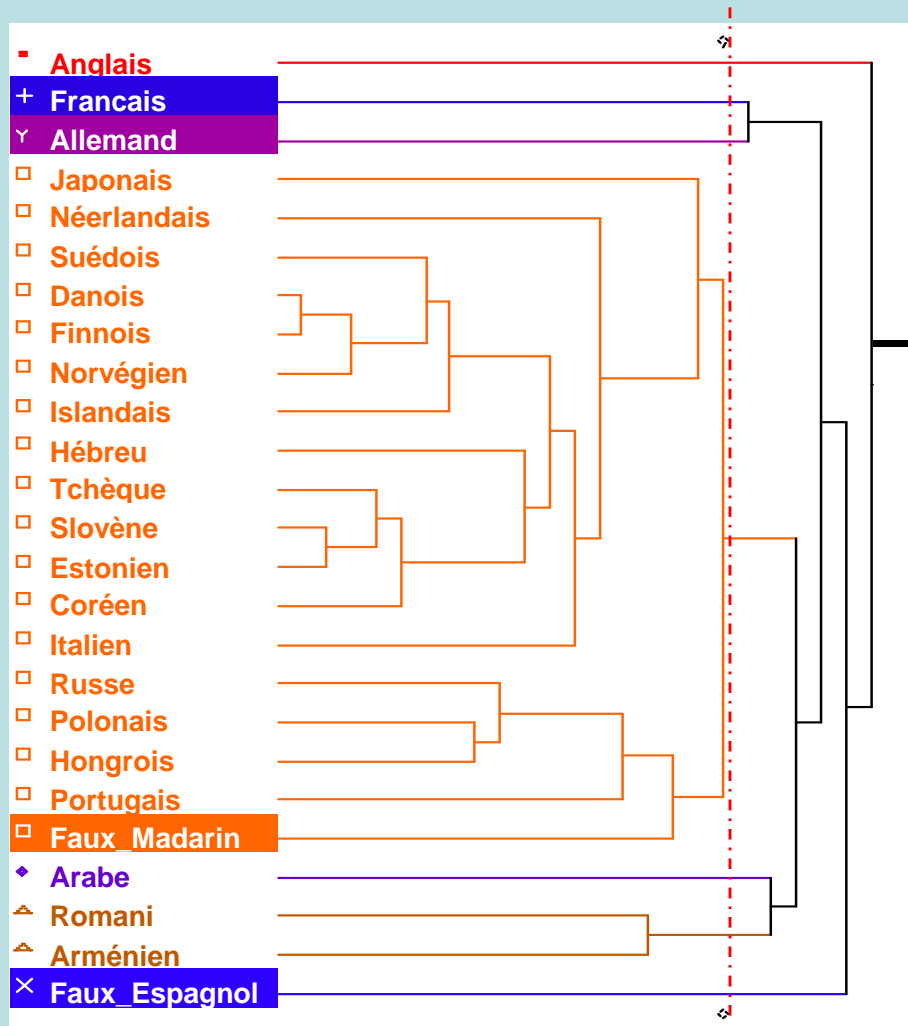
Tous les autres facteurs sont inchangés

Ajoutons deux fausses langues



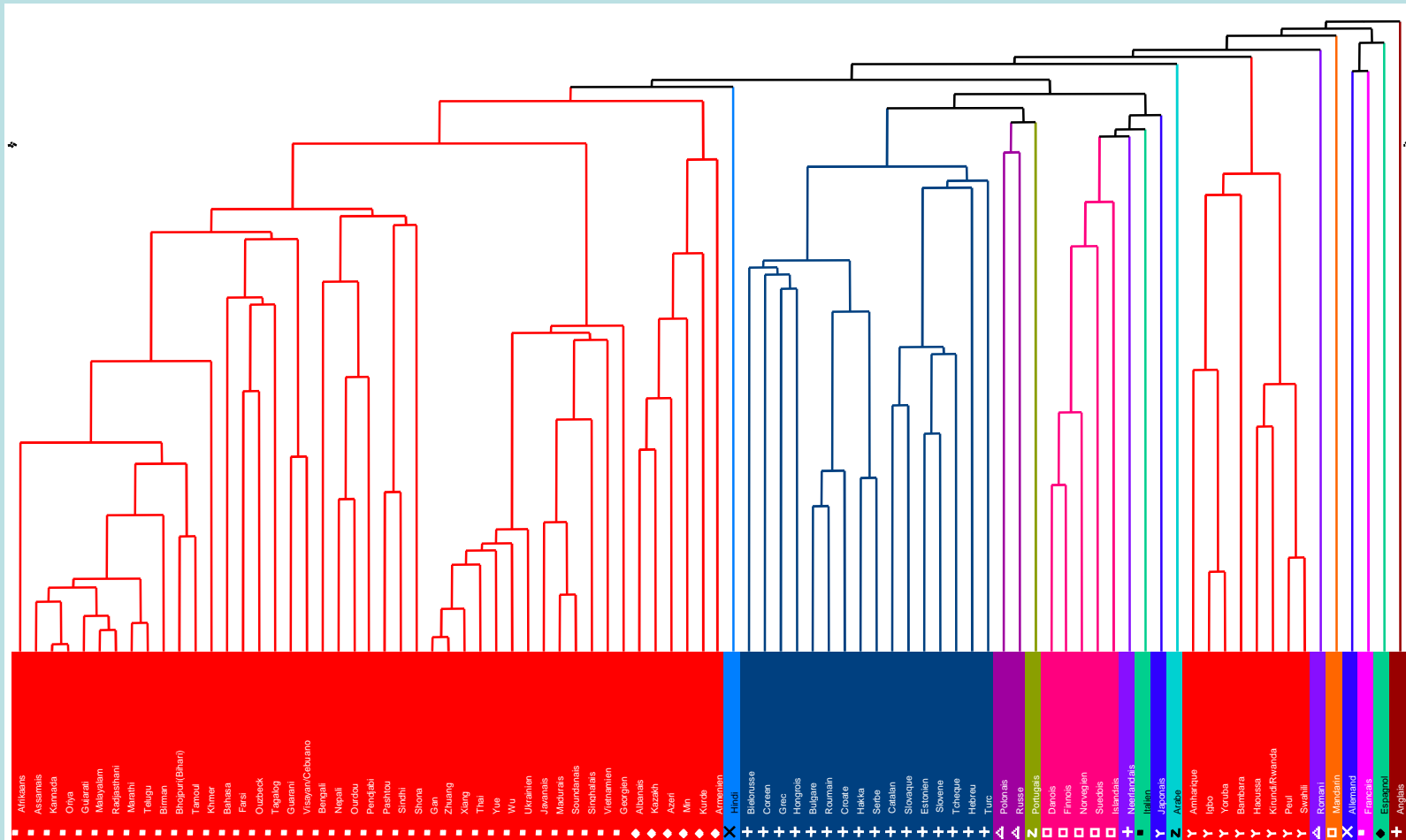
- L'espagnol et l'espagnol modifié sont groupés dans un doubleton, la modification du nombre de pays dans lesquels l'espagnol est langue officielle ne suffit pas à les séparer.
- Plusieurs facteurs assurent le poids de l'espagnol : 10 prix Nobel, 300M de locuteurs, etc ...
- Le mandarin reste un singleton.
- Le mandarin modifié est inclus dans le cluster le plus important : **seul son grand nombre de locuteurs fait du mandarin est langue « de poids »**.
- La composition des autres clusters n'est pas modifiée.

Supprimons les deux langues « vraies »



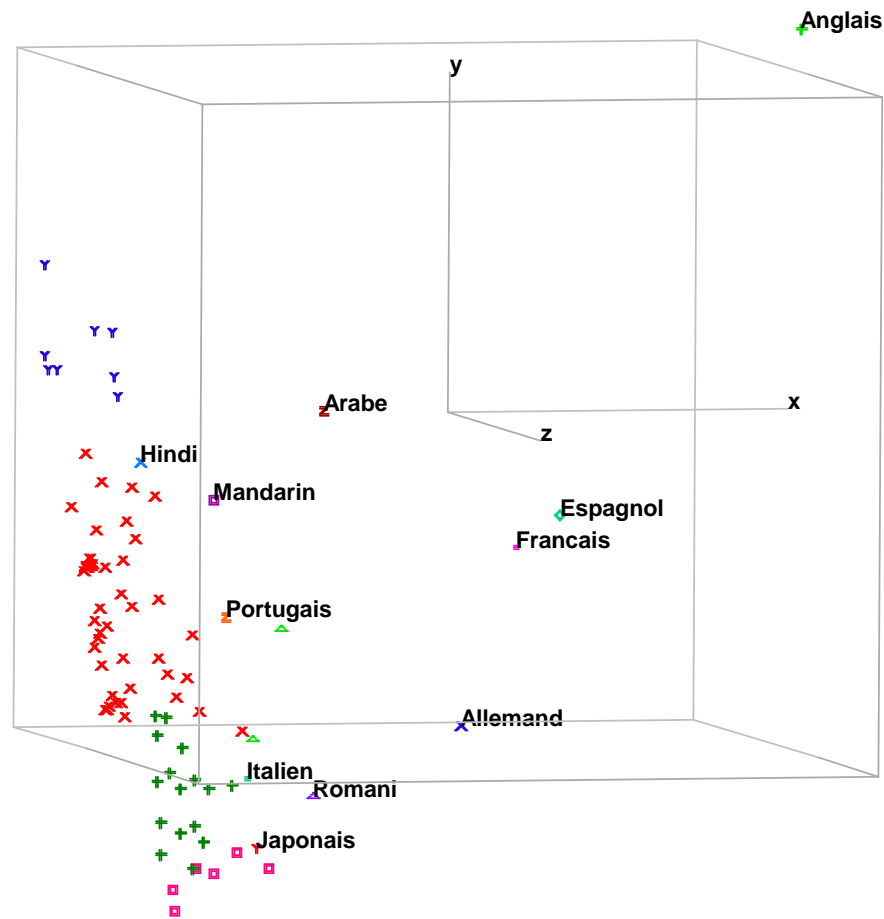
- Et si nous ne conservons que les « fausses » langues :
- le « faux » espagnol reste un singleton
- tandis que le « faux » mandarin se regroupe avec le portugais, le hongrois, etc...
- Notez aussi que le français et l'allemand se séparent, ils constituaient le doubleton le moins homogène.

88 langues, 10 facteurs, 17 clusters



Aix-en-Provence 27-28 septembre
2007

Dans un espace à trois dimensions



Utilité de la méthode

L'analyse systématique du comportement des langues modifiées permet donc de :

- Sur le plan théorique, comprendre pourquoi ces langues apparaissent uniques ou se regroupent et mettre en évidence l'importance de tel ou tel facteur
- Sur le plan pratique proposer d'éventuelles interventions sur les situations linguistiques

Pour finir..

Il s'agit donc de la
préfiguration d'un projet en
cours de réalisation.

Aix-en-Provence 27-28 septembre
2007

Comment sera-t-il présenté?

- Sur Internet (en accès libre)
- Avec la possibilité d'avoir accès à nos sources et à leur traitement
- Et donc de les critiquer et d'avoir un débat scientifique

Quelle utilité?

- 1. Mettre au point un observatoire du « poids » des langues
- 2. Qui constituera une aide à la décision en matière de politique linguistique
- 3. Et un lieu d'échange et de débats sur ces thèmes

Classement de 88 langues

(Somme des valeurs)

1	Anglais	23	Coréen	45	Pendjabi	67	Ukrainien
2	Français	24	Arménien	46	Swahili	68	Khmer
3	Espagnol	25	Estonien	47	Azéris	69	Afrikaans
4	Allemand	26	Turc	48	Bengali	70	Soundanais
5	Japonais	27	Grec	49	Min	71	Géorgien
6	Néerlandais	28	Catalan	50	Hakka	72	Bhojpuri/Bihari
7	Arabe	29	Slovaque	51	Visayan/Cébuano	73	Xiang
8	Suédois	30	Croate	52	Ourdou	74	Madurais
9	Italien	31	Kurde	53	Sindhi	75	Gan
10	Danois	32	Biélorusse	54	Vietnamien	76	Zhuang
11	Islandais	33	Albanais	55	Bambara	77	Gujarati
12	Finnois	34	Roumain	56	Ouzbèque	78	Telugu
13	Romani	35	Kirundi/Rwanda	57	Yoruba	79	Marathi
14	Russe	36	Tagalog	58	Pashtou	80	Malayalam
15	Polonais	37	Hindi	59	Tamoul	81	Singhalais
16	Portugais	38	Bahasa	60	Nepali	82	Radjasthani
17	Norvégien	39	Bulgare	61	Guarani	83	Kannada
18	Mandarin	40	Farsi	62	Igbo	84	Oriya
19	Hébreu	41	Serbe	63	Thai	85	Amharique
20	Tchèque	42	Kazakh	64	Javanais	86	Assamais
21	Slovène	43	Haoussa	65	Yue	87	Shona
22	Hongrois	44	Peul	66	Wu	88	Birman

Aix-en-Provence 27-28 septembre
2007

Classement de 88 langues

(Somme des rangs)

1	Anglais	23	Bahasa	(15)	45	Biélorusse	(- 13)	67	Oriya	(17)	
2	Français	24	Farsi	(16)	46	Thai	(17)	68	Pashtou	(- 10)	
3	Espagnol	25	Norvégien	(- 8)	47	Kazakh	(- 5)	69	Kirundi/Rwanda	(- 34)	
4	Allemand	26	Grec	(1)	48	Kurde	(- 17)	70	Wu	(- 4)	
5	Néerlandais	(1)	27	Croate	(3)	49	Ouzbèque	(7)	71	Soundanais	(- 1)
6	Russe	(8)	28	Slovène	(- 7)	50	Télugu	(28)	72	Assamais	(14)
7	Portugais	(9)	29	Slovaque		51	Marathi	(28)	73	Yoruba	(- 16)
8	Italien	(1)	30	Bengali	(18)	52	Malayalam	(28)	74	Bhojpuri(Bihari)	(- 2)
9	Arabe	(- 2)	31	Islandais	(- 20)	53	Romani	(- 40)	75	Radjasthani	(7)
10	Suédois	(- 2)	32	Albanais	(1)	54	Géorgien	(17)	76	Khmer	(- 8)
11	Japonais	(- 6)	33	Bulgare	(6)	55	Visayan	(- 4)	77	Peul	(- 33)
12	Polonais	(3)	34	Estonien	(- 9)	56	Sindhi	(- 3)	78	Amharique	(7)
13	Danois	(- 3)	35	Catalan	(- 7)	57	Gujarati	(20)	79	Xiang	(- 6)
14	Hongrois	(8)	36	Hindi	(1)	58	Swahili	(- 12)	80	Birman	(8)
15	Coréen	(8)	37	Arménien	(- 13)	59	Javanais	(5)	81	Singhalais	
16	Turc	(10)	38	Tamoul	(21)	60	Haoussa	(- 17)	82	Bambara	(- 27)
17	Finnois	(- 5)	39	Ourdou	(13)	61	Népal	(- 1)	83	Shona	(4)
18	Tchèque	(2)	40	Azéris	(7)	62	Kannada	(21)	84	Guarani	(- 23)
19	Hébreu		41	Vietnamien	(13)	63	Min	(- 14)	85	Gan	(- 10)
20	Serbe	(21)	42	Tagalog	(- 6)	64	Yue	(1)	86	Zhuang	(- 10)
21	Mandarin	(- 3)	43	Ukrainien	(24)	65	Hakka	(- 15)	87	Igbo	(- 25)
22	Roumain	(12)	44	Pendjabi	(1)	66	Afrikaans	(3)	88	Madurais	(- 14)

Aix-en-Provence 27-28 septembre
2007

!:

A suivre, donc...

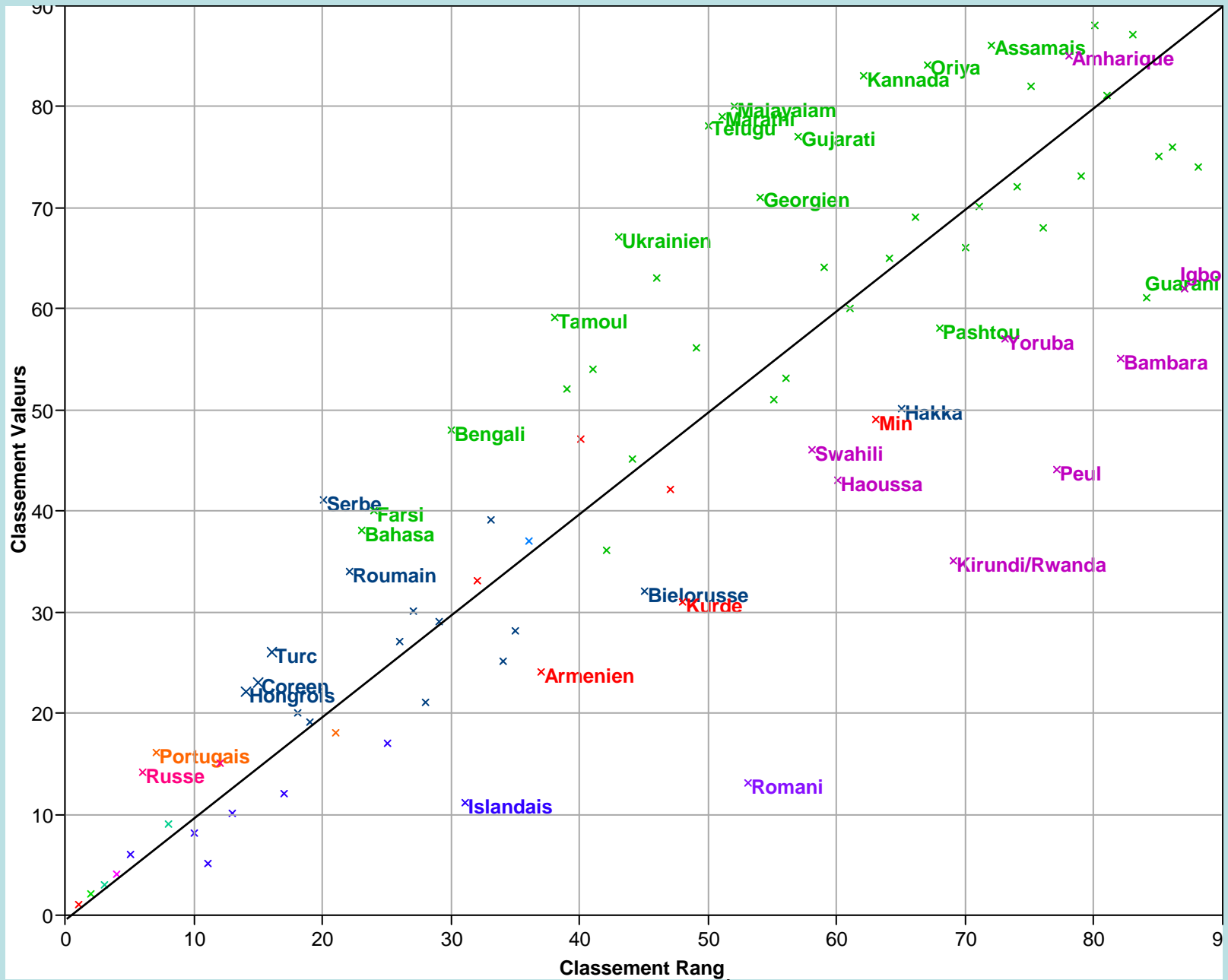
Et merci!

Aix-en-Provence 27-28 septembre
2007

That's all folks



Aix-en-Provence 27-28 septembre
2007



2007