Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ?
Application au défi DEFT 2007

Juan Manuel Torres-Moreno
Marc El-Bèze, Fréderic Béchet, Nathalie Camelin

3 juillet 2007
Laboratoire Informatique d'Avignon/UPRES 931
Université d'Avignon et des Pays de Vaucluse

Corpus disponibles

- **Critiques** (films, livres, spectacles, bandes dessinées)
  - 3 460 critiques {0,1,2}
- **Jeux vidéo**
  - 4 231 tests {0,1,2}
- **Relectures**
  - 1 484 relectures d'articles scientifiques {0,1,2}
- **Débats**
  - 28 832 interventions parlementaires {0,1}
- 60% des corpus en **apprentissage** ; 40% **test**

Evaluation

$$Fscore(\beta) = \frac{(\beta^2 + 1) \times \langle Précision \rangle \times \langle Rappel \rangle}{\beta^2 \times \langle Précision \rangle + \langle Rappel \rangle}$$

$$\langle Précision \rangle = \frac{\sum_{i=1}^{n} Précision_i}{n} \; ; \; \langle Rappel \rangle = \frac{\sum_{i=1}^{n} Rappel_i}{n}$$

$$Précision_i = \frac{\{\text{Nb de documents correctement attribués à la classe } i\}}{\{\text{Nb de documents attribués à la classe } i\}}$$

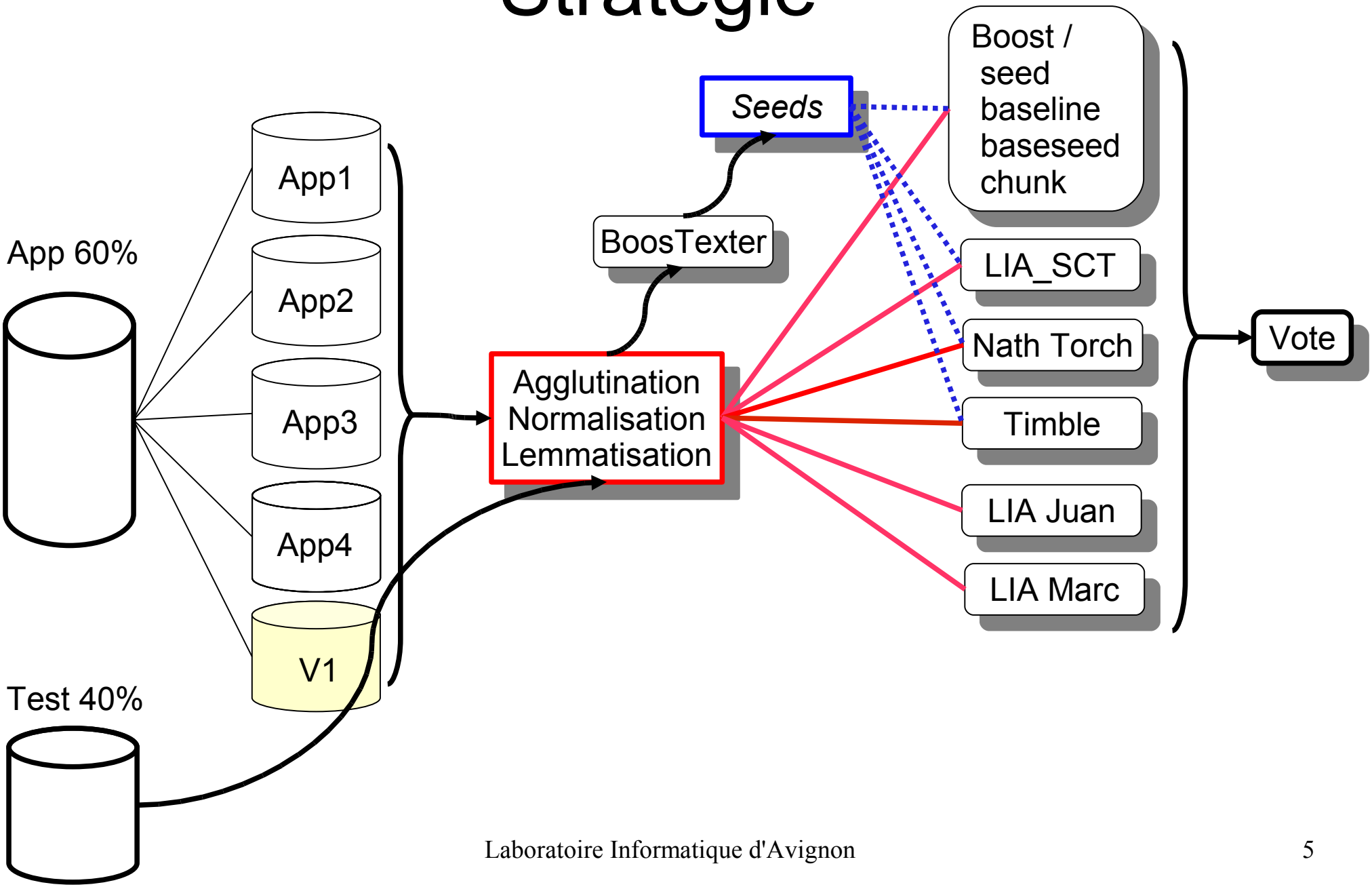
$$Rappel_i = \frac{\{\text{Nb de documents correctement attribués à la classe } i\}}{\{\text{Nb de documents appartennant à la classe } i\}}$$

Indice de confiance ponderée: probabilité d'un document d’appartenir à une classe d’opinion donnée

Représentation de textes

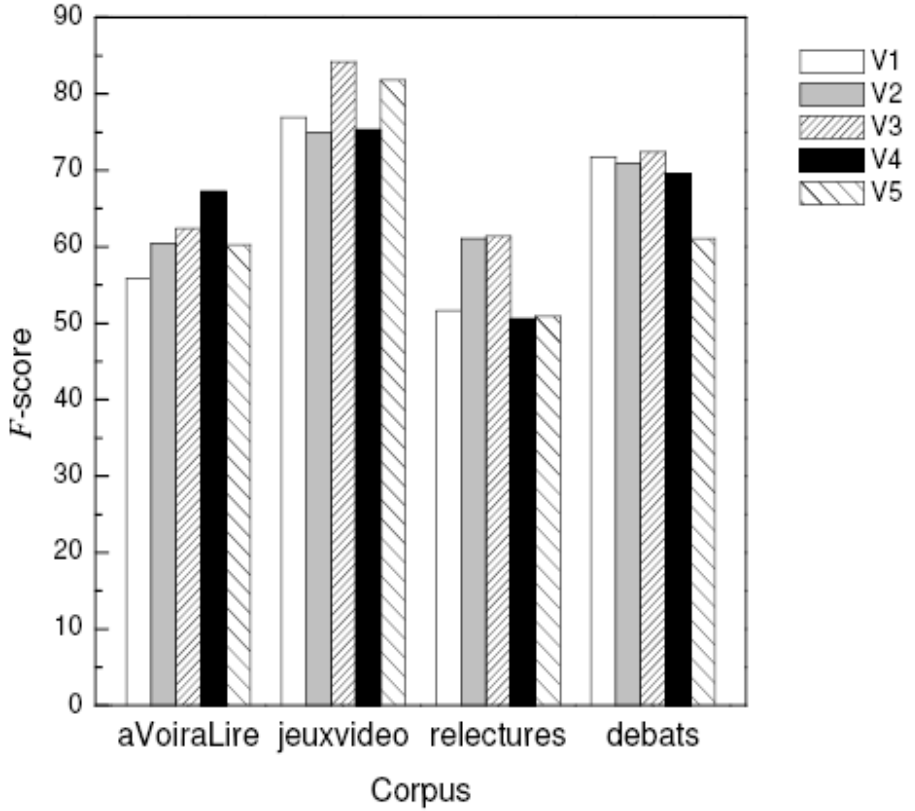
- Mots, POS, *n*-grammes, *n*-lemmes
- Etiquette supplémentaire: le *seed* *(Wilson et al., 2005)*
  - Polarité d'opinions
  - Liste manuelle
  - BoosTexter : *seed* remplacé par son lemme : lexique de 2000 *seeds*
- Normalisation/Agglutination/Lemmatisation
  - Avec ou sans

Stratégie


App 60%
App1
App2
App3
App4
V1
Test 40%
Seeds
BoosTexter
Agglutination
Normalisation
Lemmatisation
Boost /
seed
baseline
baseseed
chunk
LIA_SCT
Nath Torch
Timble
LIA Juan
LIA Marc
Vote

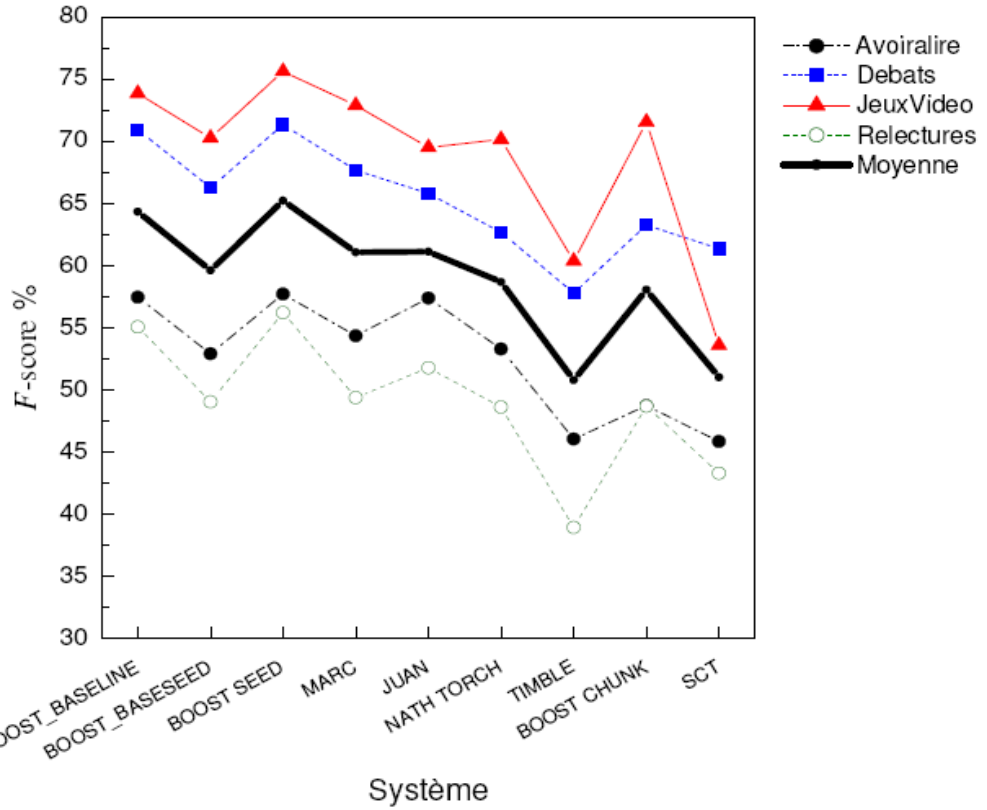
Corpus de validation

Corpus	Précision	Rappel	F-score	Correctes	Total
aVoiraLire (V)	0,6419	0,5678	0,6026	1 385	2 074
jeuxvideo (V)	0,8005	0,7730	0,7865	2 005	2 537
relectures (V)	0,5586	0,5452	0,5518	510	881
debats (V)	0,7265	0,7079	0,7171	12 761	17 299

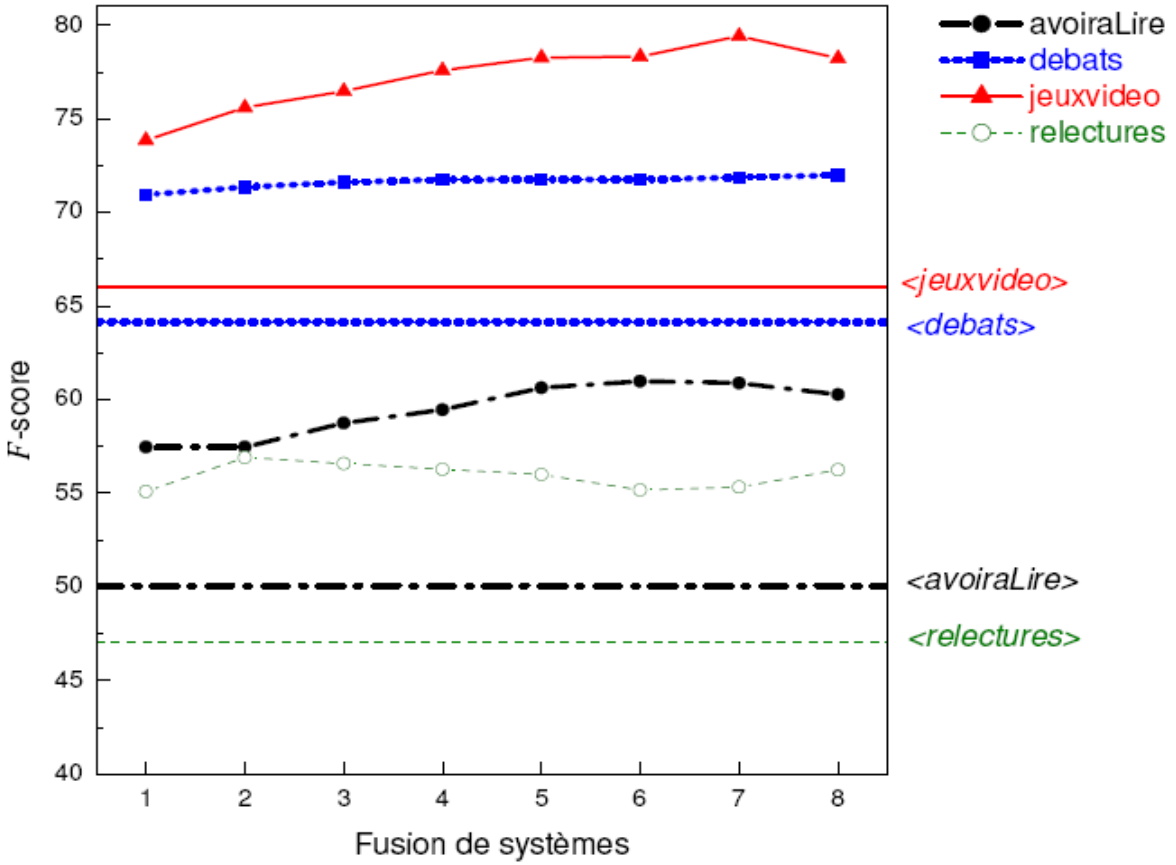


Corpus de test (1)

Corpus	Précision	Rappel	F-score	Correctes	Total
aVoiraLire (T)	0,6540	0,5590	0,6028	931	1 386
jeuxvideo (T)	0,8114	0,7555	0,7824	1 333	1 694
relectures (T)	0,5689	0,5565	0,5626	353	603
debats (T)	0,7307	0,7096	0,7200	8 403	11 533

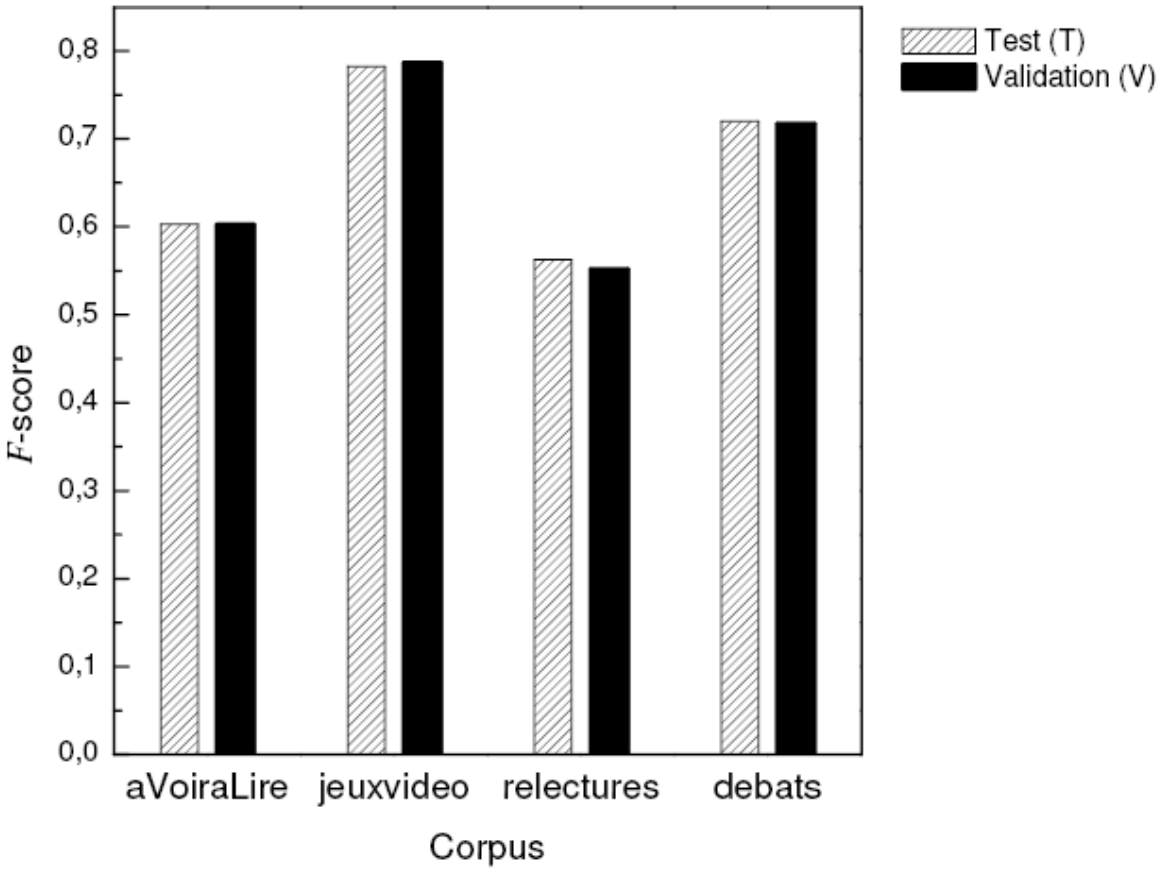


Corpus de test (2)



1 BOOST_BASELINE
2 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED
3 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + MARC
4 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN
5 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH
6 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH + TIMBLE
7 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH + TIMBLE + BOOST_CHUNK
8 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH + TIMBLE + BOOST_CHUNK + SCT

Corpus de test (3)



Discussion

3 :2 relectures
L'idée d'appliquer les méthodes de classification pour définir des classes homogènes de pages web est assez originale par contre, la méthodologie appliquée est classique. Je recommande donc un « weak accept » pour cet article.

Classé **1** par le système (*accepté avec des modifications majeures*).
De la lecture directe on pourrait en déduire que la classe est **1**...
mais la référence est **2** (*accepté*)

3.6 relectures
Article trop court pour pouvoir être jugé. Je suggère de le mettre en POster si cela est prévu.

Document trop court : les systèmes l'ont rejetté, mais il a été accepté

3 :9 relectures

Question : comment est construit le réseau bayesien ? Un peu bref ici... Remarques de forme : page 2, 4ème ligne, « comprend » 5ème ligne : "annotées" ou "annoté" page 3 : revoir la phrase confuse précédant le tableau dernière ligne, répétition de "permet" page 5 : 7ème ligne accorder "diagnostiqué" et "visé" avec "états" ou avec "connaissances"

Rejetté mais il a été accepté... sur des remarques de forme uniquement !

aVoiraLire 1 :10

Depuis trente-six ans, chaque nouvelle production de David Bowie est un événement. Heathen , ne fait pas exception à cette règle. On reconnaît instantanément la patte de son vieux compère Tony Visconti. La voix de Bowie est mise en avant. Agréable surprise, surtout qu'elle n'a rien perdu depuis ses débuts. Là, commence le voyage. Ambiance, mélange dosé des instru[illegible] Dès l'ouverture de l'album avec Sunday , un sentiment étrange nous envahit. Comme [illegible]Bowie venait de rentrer d'un voyage expérimental au coeur même de la musique. [illegible]tour aux sources. L'ensemble du disque est rythmé par cette pulsation dont le du[illegible] le secret. Le tout saupoudré de quelques pincées d'électronique. Le groupe est [illegible]uit au minimum. Outre Bowie en chef d'orchestre et Visconti, David Torn ponctu[illegible]es compositions de ses guitares aventureuses et Matt Chamberlain apporte de l'âme à la rythmique. Un quatuor à cordes fait une apparition, comme Pete Townshend (Th[illegible]o) ou Dave Grohl (ex-batteur de Nirvana). Avec trois reprises réarrangées et neuf compositions originales, le 25e album de Bowie est à l'image d'une cohérence artistique retrouvée.

Conclusions

- L'union fait l'*intelligence* : combinaison de plusieurs méthodes numériques/probabilistes
- Indépendance de la langue
- Indépendance du contexte
- Résultats honorables
- Adaptable à d'autres problématiques