Présentation de la campagne ESTER

tâches, corpus et ressources

Guillaume Gravier

Association Francophone de la Communication Parlée

Plan

1. Campagne d’évaluation
2. Description du corpus
3. Traitement(s) automatique(s)
4. Conclusion

Plan

1. Campagne d’évaluation
   - organisation
   - tâches
   - résultats
2. Description du corpus
3. Traitement(s) automatique(s)
4. Conclusion

Organisation de la campagne ESTER

projet Technolangue EVALDA

- **Association Francophone de la Communication Parlée**
  - animation scientifique
  - définition du plan d'évaluation
  - valorisation scientifique
- **Centre d'Expertise Parisien de la DGA**
  - définition du plan d'évalation
  - production de ressources
  - validation des résultats
- **ELDA**
  - production de ressources
  - diffusion des ressources

Organisation de la campagne ESTER

- Calendrier
  - démarrage à l'été 2003
  - première phase de test à blanc en janvier 2004
  - test officiel en janvier 2005
  - atelier de clôture en mars 2005
- Participants
  - une dizaine de laboratoires
  - principalement académiques
  - l'ensemble des acteurs de la communauté traitement automatique

Tâches évaluées

Traitement automatique des informations radiodiffusées

(S) Segmentation

(T) Transcription orthographique / lexicale

(E) Extraction d'information

Tâches : segmentation

Objectif : produire un découpage suivant un critère donné

(SES) Segmentation en événements sonores
- détection des zones contenant de la parole
- détection des zones contenant de la musique

(SRL) Segmentation et regroupement en locuteurs
- détection les changements de locuteurs
- grouper les parties du documents correspondant à un même locuteur
- pas de connaissance a priori

(SVL) Suivi de locuteurs
- détection des zones correspondants à un locuteur donnée
- liste de locuteurs cibles, connus a priori

Tâches : transcription

Objectif : produire une transcription lexicale

- évaluation sur la base du “*mot*”
- **transcription normalisée** : pas de ponctuation, pas de casse, etc...

  [...] <s> si vous êtes euh adeptes de l’ invective et du bras d’ honneur en voiture en abstenait vous au moins pour aujourd’hui <s> c’ est la troisième journée nationale de la courtoisie au volant <s> un rendez vous qui tombe à la veille du week-end de pâques l’ un des plus meurtriers de l’ année <s> bison futé voit rouge dans le sens des départs [...] `[play]`
- deux catégories de systèmes
  1. temps réel : capable de traiter une heure de données en approximativement une heure
  2. non contraint : pas de limite sur le temps de calcul (entre 10 et 20 fois temps réel)

Tâches : extraction d'informations

Objectif : extraire des informations de haut-niveau sémantique

- détection des entités nommées : noms propres, noms de lieu, dates, mesures, etc...
- *segmentation et détection thématique*
- *question-réponse*

Seule une version expérimentale de la tâche EN a été réalisée

Résultats

laboratoire	TRS	TTR	SES		SRL	SVL	EN	
			par.	mus.			ref	asr
CLIPS	40,7				27,2			
ENST/TSI	45,4					47,0		
FT R&D			99,1					
IRISA-ENST/INF	35,4		98,9	33,7		**84,3**		
IRIT	61,9	70,4	98,8	52,7				
LIA	26,7	36,3	**99,2**	**54,8**	19,2	66,0	**34,1**	**57,4**
LIMSI	**11,9**				**11,5**			
LIPN							37,0	
LIUM	23,6		97,4		16,9		39,7	61,2
LORIA	27,6	37,4	97,5					
Univ. Balamand			95,1	26,2				
Vecsys Research		**16,9**						

Plan

1. Campagne d’évaluation
2. Description du corpus
3. Traitement(s) automatique(s)
4. Conclusion

Plan

1. Campagne d’évaluation
2. Description du corpus
   - description générale
   - principes d’annotation
   - phénomènes annotés
3. Traitement(s) automatique(s)
4. Conclusion

Description générale

- Quatre sources principales : France Inter, France Info, Radio France International et Radio Télévision Marocaine
- Tranches horaires d'informations incluant
  - information studio : titres et développement des titres
  - reportages
  - annonces et publicités
  - sonaux, génériques et intermèdes musicaux
- Type de parole
  - planifiée (principalement)
  - plus ou moins spontannée : interviews, reportages, etc.

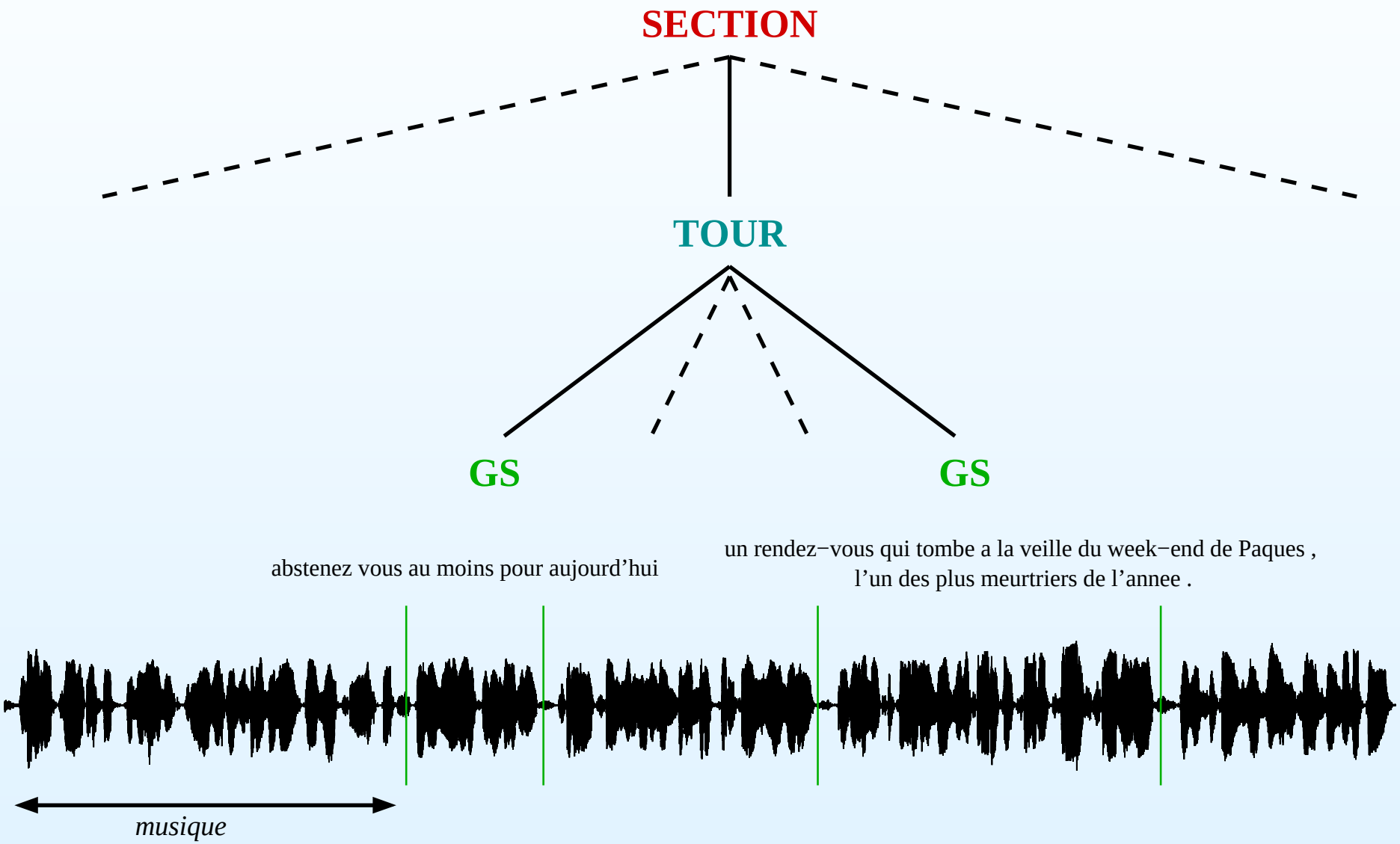
Volume de données

Corpus	Inter	Info	RFI	RTM	Cult.	Clas.	Total
apprentissage	33h	8h	23h	21h	–	–	85h
dévellopement	2h	2h	2h	2h	–	–	8h
test	2h	2h	2h	2h	1h	1h	10h
Total	37h	12h	27h	25h	1h	1h	$\simeq$ 100h

- Inter : 1998-1999, 2003
- RFI : 2000, 2003, 2004
- Les autres : 2003, 2004
- Corpus de test : 2004

Grand volume de données non transcrites.

Annotation : Principe



Annotation : Sections

une section = une entité journalistique

- report
  - indication thématique associée, par exemple
    - les titres du journal
    - la météo
    - spéciale 21 avril
    - à votre santé
  - mais pas très normalisé
- filler
- nontrans

Annotation : Tours de parole

un tour de parole = une entité locuteur

- zones de double parole spécifiées
- locuteurs globaux ou locaux (variabilite a long terme)
- attributs
  - sexe
  - natif ou pas (attention à RTM)
  - mode spontané ou planifié
  - canal de transmision (studio ou téléphone)

Annotation : Groupes de souffles

transcription synchronisée sur les groupes de souffles

- ponctuation présente mais pas de majuscule en début de phrase
- indications de phénomènes non linguistiques
  - inspirations, bruits de bouche, etc...
  - top sonores, jingles, indicatifs, etc...
- marques de phénomènes linguistiques
  - mots prononcés dans une langue etrangère
  - mots tronqués, reprises, etc...
  - prononciations non standards

Annotation : à l'écoute!

Écoutons et voyons...

Phénomènes linguistiques

- Mots tronqués, reprises et hésitations

[play] *euh nous avons montré , chose fa(cile) () de façon très claire , qu'il s'agissait d'un mariage , d'un sous mariage ,*

[play] *combattant dep(uis) () des idées euh fascistes et racistes depuis très , très longtemps*

[play] *eh ben , (il) y a que le coup d'état euh pour le remplacer : c'est bien clair et c'est à quoi Bruno Mégret s'em(ploie) () s'emploie*

[play] *par ailleurs euh , euh , y faut toujours bouger euh les responsables*

Phénomènes linguistiques

- Prononciations non standards

[play] *ça donne une majorité* ***hétéroclite** euh , hétéroclite , pardon de Julien Dray , Dominique Strauss-Kahn pour résumer en passant*

[play] *malgré la dégradation de l'environnement* ***économique** économique , et le nouveau coup de tabac sur les marchés financiers .*

[play] *sur le rôle du tribunal* ***pénal** pénal international*

- Sigles et acronymes

[play] *la résolution 687 de l'_ONU à la fin de la guerre du Golfe prévoyait en Irak*

[play] *les frappes contre l'Irak violent la charte de l'!ONU*

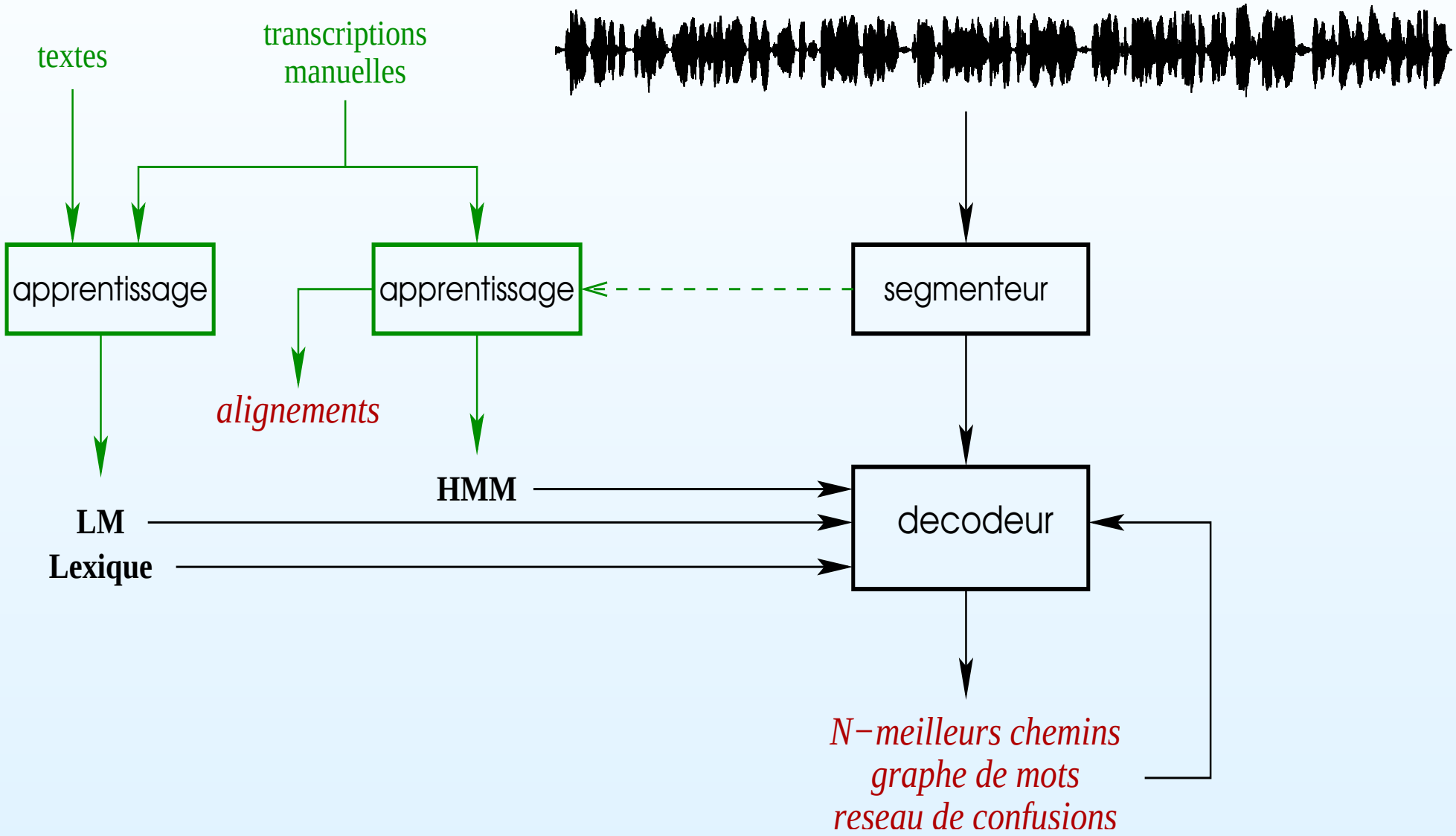
Plan

1. Campagne d’évaluation
2. Description du corpus
3. Traitement(s) automatique(s)
4. Conclusion

Plan

1. Campagne d’évaluation
2. Description du corpus
3. Traitement(s) automatique(s)
   - architecture
   - produits dérivés
   - quelles erreurs
4. Conclusion

Architecture d'un système de RAP



Ressources d'un système de RAP

- Lexique
  - une entrée = une forme graphique
  - variantes de prononciations, éventuellement probabilisés

affiches	a f i S
affiches	a f i S z
affiches	a f i S @ z
affichette	a f i S E t
affichettes	a f i S E d z
affichettes	a f i S E t
affichettes	a f i S E t @ z
affichettes	a f i S E t z

  - pas de contraintes phonétiques ou linguistiques
    (par exemple sur les liaisons)

Ressources d'un système de RAP

- Modèle de langage
  - modèles portant sur les formes graphiques
  - modèles à courte portée (n-gram avec $n < 5$)
  - pas de contraintes linguistiques
- Modèles acoustiques
  - modèles de phones en contextes

    affiches *_a_f a_f_i f_i_S i_S_*

    affiches *_a_f a_f_i f_i_S i_S_z S_z_*

    affiches *_a_f a_f_i f_i_S i_S_@ S_@_z @_z_*
  - modèles de Markov cachés

Alignements phonétiques

- choix des variantes de prononciations
  - élisions
  - liaisons
  - assimilations
  - etc.
- alignement phonétique
  - à voir...

Sortie de systèmes

- transcription

`[play]` <s> c' est ce qu' a déclaré le président pakistanais pervez moucharraf a pas dans une conférence de presse </s>

- N-meilleurs transcriptions

  <s> c' est ce qu' a déclaré le président pakistanais pervez moucharraf a pas dans une conférence de presse </s>

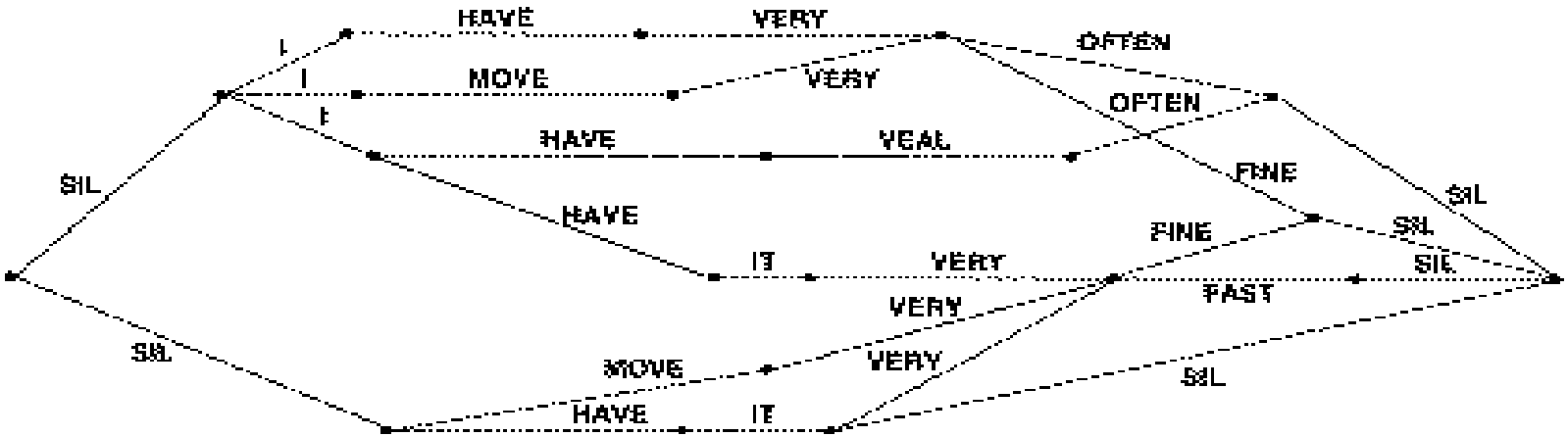
  <s> c' est ce qu' a déclaré le président pakistanais pervez moucharraf a dans une conférence de presse </s>

  <s> c' est ce qu' a déclaré le président pakistanais pervez moucharraf est dans une conférence de presse </s>

- graphes de mots
- réseaux de confusion

Sortie de systèmes

(a) Input lattice ("SIL" marks pauses)



(b) Multiple alignment ("-" marks deletions)

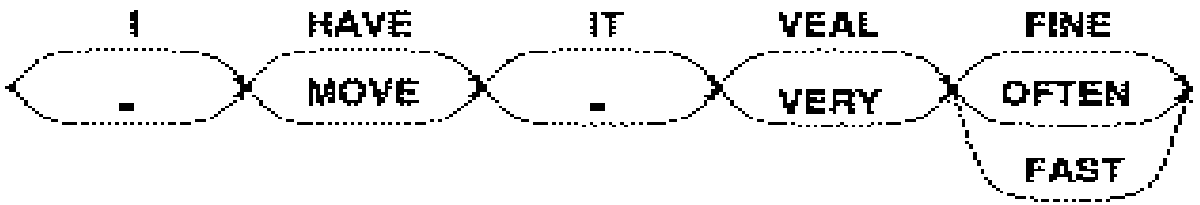


Figure 3.1: From lattices to multiple alignments

Erreurs des systèmes

- *dérapage* liés à des conditions particulières : acccents, bruits de fond, etc.
- fautes d'orthographes, particulièrement accord
- mauvaise reconnaissance des mots courts : *le*, *la*, etc...
- trop de variantes de prononciation = erreurs de confusion
- pas assez de variantes de prononciation = erreurs aussi!

Voyons voir ça...

En guise de conclusion...

- Connaissances utilisées
  - un peu de phonétique dans le lexique
  - un peu de linguistique dans le modèle de langage
- Connaissances non utilisées
  - prosodie
    - pour la segmentation en groupe de souffle
    - pour la (re)segmentation en phrase
  - Prononciations
    - adaptation des variantes prononciations
    - contraintes sur les prononciations
  - et beaucoup d'autres encore ...