Reconnaissance d’expressions polylexicales verbales sur la transcription de la parole

Nicolas ZAMPIERI (LIS) – Carlos Ramisch (LIS) – Géraldine Damnati (Orange Labs)

M2 IAAA - 2019

Sommaire

- Description de la tâche
- Expériences
- Résultats
- Perspectives

- Qu’est-ce qu’une expression polylexicale ?
  - Casser du sucre sur le dos (VID)
  - La musique adoucit les mœurs (VID)
  - Faire une présentation (LVC)
  - S’apercevoir (IRV)

Description de la tâche

- Tâche :
  - Reconnaître automatiquement les expressions.
    - David fait une présentation.
    - Une présentation est faite par David.
    - David a fait une présentation et un hommage en même temps.
    - David se fait des idées

Description de la tâche

- Tâche :
  - Reconnaître automatiquement les expressions.
    - David fait une présentation.
    - Une présentation est faite par David.
    - David a fait une présentation et un hommage en même temps.
    - David se fait des idées.

Description de la tâche

- Pourquoi est-ce une tâche importante ?

Français	Anglais (traduction littérale)	Anglais (traduction idiomatique)
je déteste **casser du sucre sur le dos** de quelqu'un.	I hate to break the sugar on someone's back.	I don't like **talking about** people **behind their back**.

Expériences

- Questions de recherches :
  - Impact de la lemmatisation (Zampieri, et al. 2019)
  - Impact des informations syntaxiques
  - Convolution sur les caractères pour les expressions non vues
  - **Performance sur le langage parlé**

Expériences

- Corpus
  - Pour l'écrit (ST) : entrainement, validation, test
  - Pour l'oral (petits bateaux) : validation
- Système
  - https://github.com/zamp13/Veyn
- Mesures d'évaluation

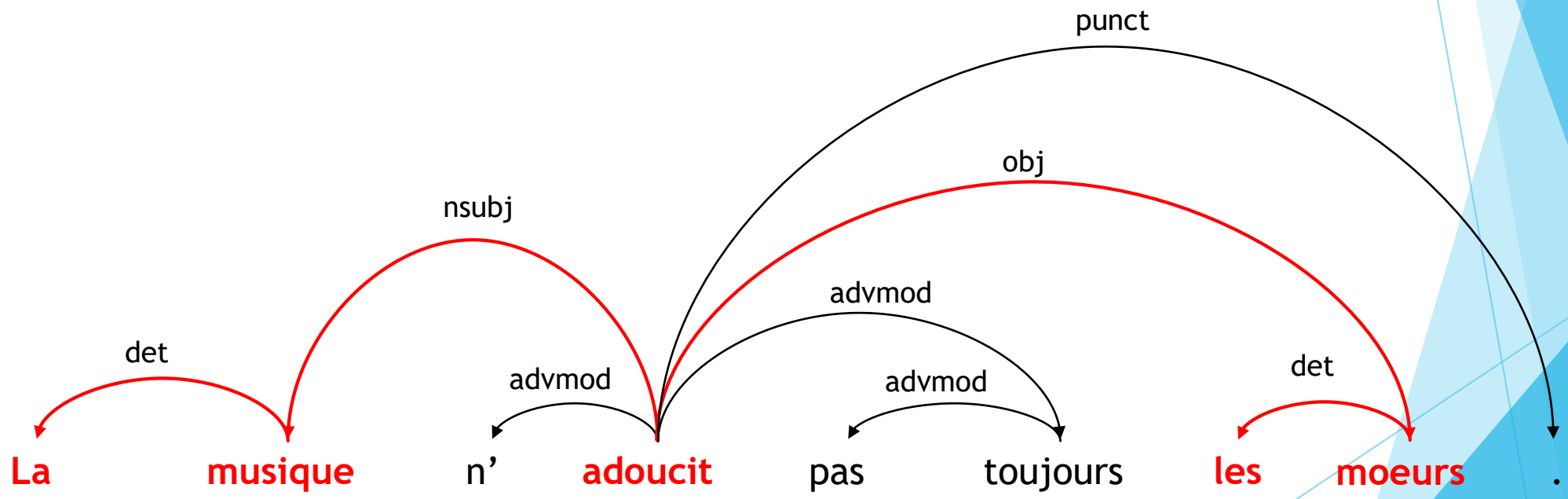
Expériences : corpus (ST)

Id	Forme	Lemme	POS	[…]	Gov	Deprel	[…]	MWE
1	La	le	DET	[…]	2	det	[…]	1:VID
2	musique	musique	NOUN	[…]	4	nsubj	[…]	1
3	n’	ne	ADV	[…]	4	advmod	[…]	*
4	adoucit	adoucir	VERB	[…]	0	root	[…]	1
5	pas	pas	ADV	[…]	6	advmod	[…]	*
6	toujours	toujours	ADV	[…]	4	advmod	[…]	*
7	les	le	DET	[…]	8	det	[…]	1
8	mœurs	mœurs	NOUN	[…]	4	obj	[…]	1
9	.	.	PUNCT	[…]	4	punct	[…]	*

- Source: corpus d’entrainement français de la campagne d’évaluation PARSEME 1.1

Expériences : corpus (ST)

- Arbre syntaxique



Expériences : corpus (ST)

- Détails des corpus utilisés

Corpus	Tokens	VMWE	Vocabulaire	
			Formes	Lemmes
EU-train	117 165	2 832	26 912	11 602
EU-dev	21 604	500	7 766	4 178
EU-test	19 038	500	7 226	3 902
FR-train	**420 762**	**4 550**	45 166	**33 928**
FR-dev	54 685	629	11 593	8 814
FR-test	38 402	498	8 160	6 052
PL-train	220 352	4 122	**48 211**	21 795
PL-dev	26 014	515	10 007	5 955
PL-test	27 661	515	10 285	6 408

Expériences : corpus (ST)

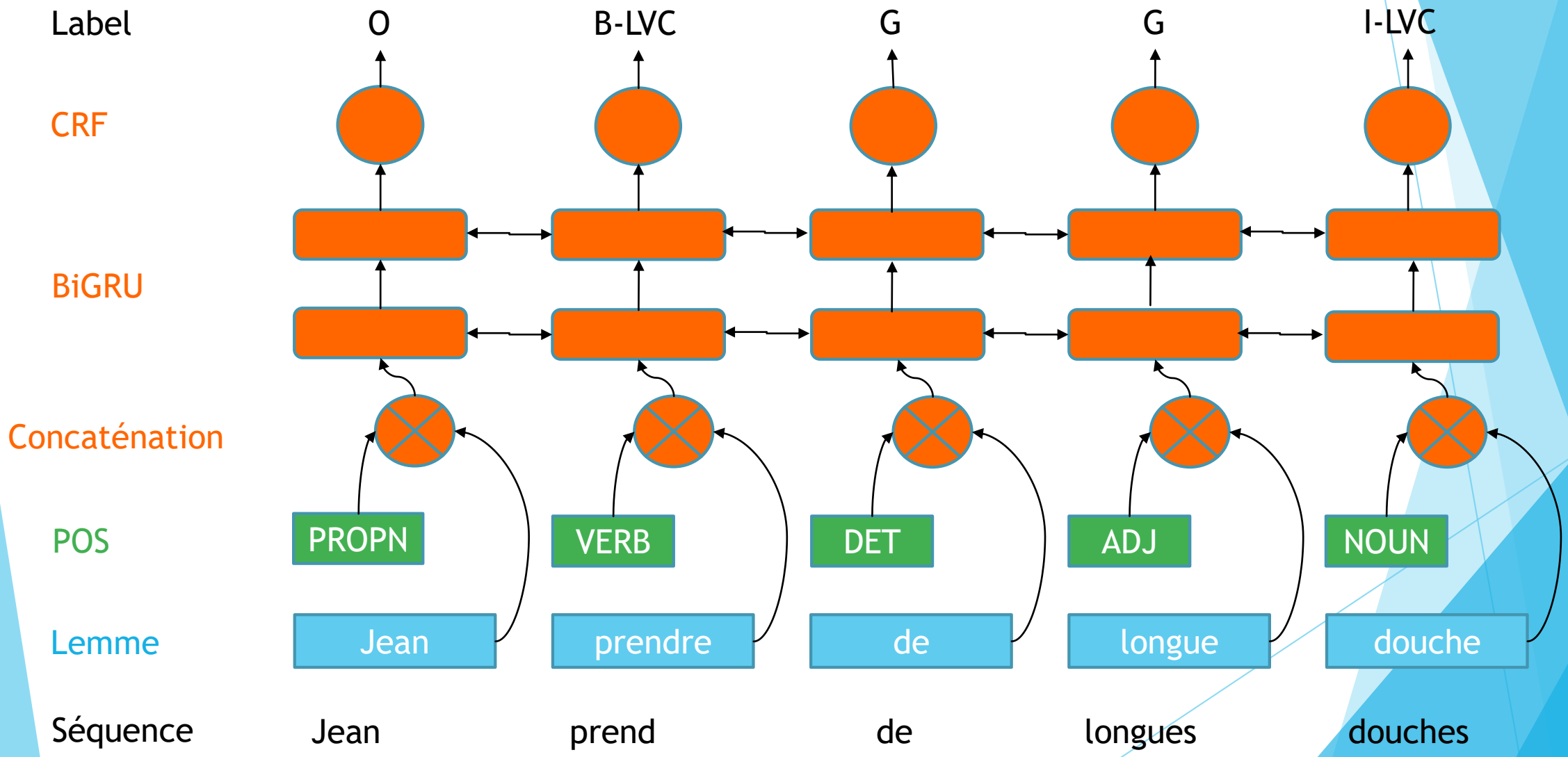
- Indice de la richesse morphologique des corpus d'entrainement
  - Ratio formes/lemmes

Corpus	Morpho
Basque	**2,32**
Français	1,33
Polonais	2,21

Expériences : corpus (petits bateaux)

- Issu d'une émission de radio
- Annotation manuelle des expressions pour le corpus de validation :
  - Annotateurs : Carlos Ramisch et Nicolas Zampieri
  - Guide d'annotation : https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/
- 760 phrases et 557 expressions
  - […] lesquelles **y a** plein d'énergie, plein de molécules organiques.
  - bonjour, Juliette, **euh**, ta **question** est bien **posée** […]

Expériences : système



Expériences : système

Label … B-LVC …

CRF

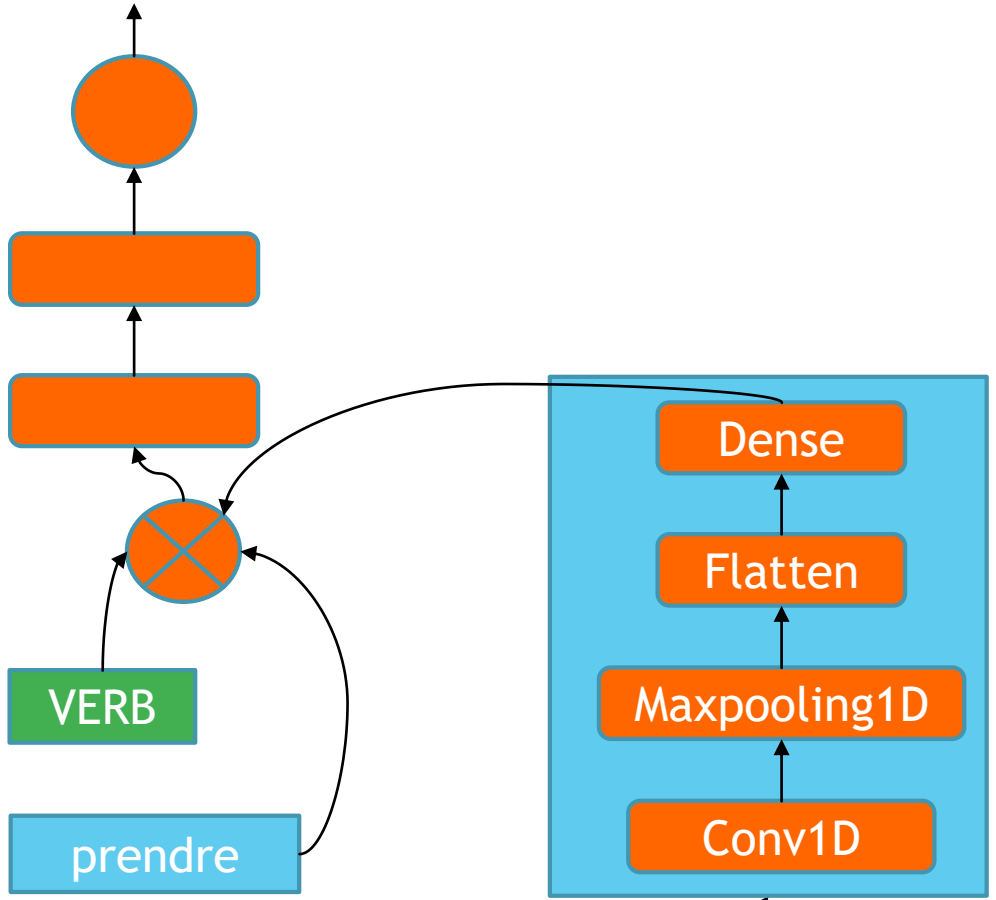
BiGRU

Concaténation

POS

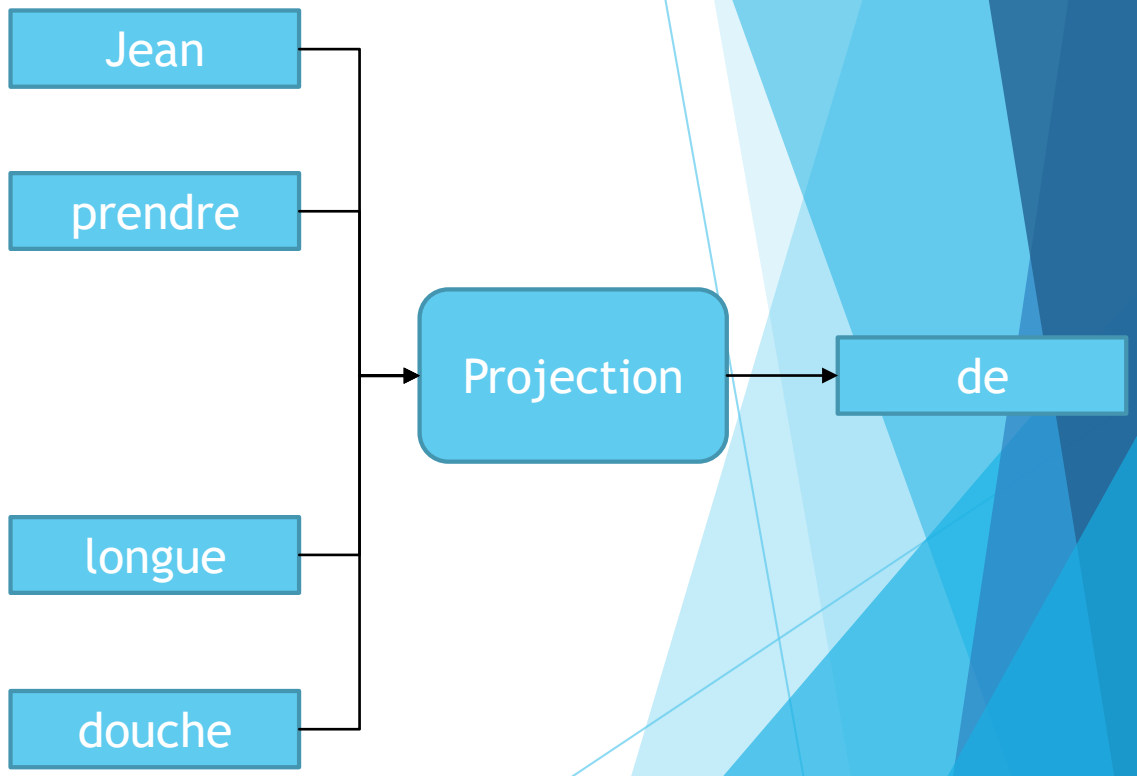
Lemme

Séquence … prend …



Expériences : système

- Représentations de mots :
  - Word2vec
    - Représentation classiques
    - Algorithme CBOW
  - FastText
    - Variante de word2vec
    - Utilisation des n-grams de caractères
    - Algorithme CBOW



Expériences : mesures d'évaluation

- F-mesures :
  - Expressions complètes (F-MWE)
  - Par token (F-TOK)
  - Expressions identiques à une expression vue (F-ID)
  - Expressions variantes à une expression vue (F-VAR)
  - Expressions non vues (F-UNS)

Résultats

- Impact de la lemmatisation

Entrées	Embeddings	Basque	Français	Polonais
		F-MWE	F-MWE	F-MWE
Forme	Word2vec	60,37	47,41	42,27
Forme	FastText	66,52	52,60	47,24
Lemme	Word2vec	53,36	53,28	57,82
Lemme	FastText	62,86	59,35	**61,49**
Forme-Lemme	Word2vec	60,56	56,11	56,80
Forme-Lemme	FastText	**69,24**	**60,41**	57,39

- **FastText est meilleure que word2vec**
- **La lemmatisation est importante pour les langues à forte et faible richesse morphologique**

Résultats

- Impact des informations syntaxiques

Entrées	Embeddings	Basque	Français	Polonais
		F-MWE	F-MWE	F-MWE
Forme-Lemme	FastText	**69,24**	60,41	57,39
Forme-Lemme-GovL	FastText	62,20	59,63	**60,83**
Forme-Lemme-Deprel	FastText	66,18	**64,40**	43,77
Forme-Lemme-GovL-Deprel	FastText	65,77	59,50	60,20

- **On ne peux pas réellement conclure pour l'utilisation des lemmes du gouverneur**
- **Pour le français, les étiquettes syntaxiques semblent importantes**

Résultats

- Convolution sur les caractères pour les expressions non vues

Entrées	Embeddings	Basque		Français		Polonais	
		F-MWE	F-UNS	F-MWE	F-UNS	F-MWE	F-UNS
Forme-Lemme	FastText-FastText	69,24	5,26	**60,41**	18,57	57,39	10,96
Forme-Lemme	CNN-word2vec	67,14	7,44	58,45	**22,34**	**64,15**	16,27
Forme-Lemme	CNN-FastText	**70,61**	**7,55**	59,61	20,99	63,42	**17,72**

- **La convolution sur les caractères de la forme permet d'accroitre l'identification des expressions non vues en entrainement**

Résultats

- Transcription de la parole

Entrées	Embeddings	Petits-bateaux (UD_Spoken)				
		F-MWE	F-TOK	F-UNS	F-VAR	F-ID
Forme-Lemme	FastText-FastText	61,03	74,10	8,49	57,44	**86,96**
Forme-Lemme-Deprel	FastText-FastText	62,26	74,48	15,49	62,31	86,88
Forme-Lemme	CNN-FastText	**62,32**	**75,03**	**17,30**	**65,74**	84,48

- **Notre système est performant sur la transcription de la parole**

Perspectives

- Recherche d’hyperparamètres
- Recherche d’une meilleure représentation des étiquettes
- Utilisation du système pour d’autres tâches

Merci pour votre attention !