Quantitative Comparative Interactional Linguistics

Laurent Prévot

Variamu 3rd Workshop, October, 1st-2nd, 2015

Interactional Linguistics

What it is?

- how people are interacting with each other through language
- the study of the linguistic structures of such interaction

Focus on

- analysis of spontaneous spoken data
- objects studied are multidimensional *(lexis, syntax and prosody,...)*
  - turn-taking, discourse particles, discourse syntactic positions, repairs, fragments, spoken language constructions

[Couper-Kuhlen and Selting, 2001]
Methods:

- Conversational Analysis
- light-weight quantitative descriptions (sometimes)

Comparative Interactional Linguistics

Contrastive Conversation Analysis [Maynard, 1990]
Studied multimodal backchannel behaviors in English and Japanese (*aizuchi*)
Says that backchannels in Japanese and English occurs in different contexts

- Corpus-based: about 2 hours of video
- Manual coding and analysis
- problem of *'equivalence'*: cannot rely on semantic equivalence through parallel data / sentences

[Clancy et al., 1996]: Mandarin, English, Japanese (25 minutes)

More Interactional? Linguistics: Discourse and semantic studies

- [Lambrecht, 1988]: SVO with lexicalized S and O is not the basic structure for spoken French
- [Traugott and Dasher, 2001]'s paths of semantic change
  - truth-conditional $\rightsquigarrow$ non-truth conditional (?)
  - content $\rightsquigarrow$ content-procedural $\rightsquigarrow$ procedural (?)
  - scope-within-proposition $\rightsquigarrow$ scope-over-proposition $\rightsquigarrow$ scope-over-discourse
  - nonsubjective $\rightsquigarrow$ subjective $\rightsquigarrow$ intersubjective

More Interactional Linguistics: Formal approaches to dialogue

[Ginzburg, 2012] accumulates example to justify

- the promotion of *tokens* (vs. *types*) as first-class citizens for grammar
- a grammar of performance
- the inclusion of a dialogue game board with public and private parts

Formalized (in an HPSG-style grammar boosted with situation semantics and expressed in TYPE THEORY WITH RECORDS) :

- short answers, clarification ellipses
- simple feedback
- disfluencies

Quantitative Comparative Interactional Linguistics

- *quantitative* requires significant amount of data (statistical significancy)
- QCIL : Approach in a systematic a data-driven way on large comparable corpora
- Existing works :
  - [Ward and Tsukahara, 2000]: Turn-taking and prosody in English and Japanese
  - [Levitan et al., 2015]: Entrainment in English, Mandarin, Spanish and Slovak
  - ...

General framework

Same situation encoded in comparable corpora

- same communicative needs
- same time pressure
- same interpersonnal relationships
- (remain interindividual variation)

Significant differences observed due to:

- linguistic / interactional structures
- socio-cultural constraints

Commonalitites / Universals ?

- At interactional level [Levinson, 2006]
- Related to findings on Broca's area of processing complex hierarchical structures [Higuchi et al., 2009]

Overall characteristics of the 'orchid' dataset

Size:

lge	dur(m)	syll	tokens	PU	DU
fr	89	23631	20233	6057	2130
tw	205	54615	37637	8563	5673

- face-to-face interaction, long conversation, without a very specific task
- recorded in good conditions

Domains:

Description	**Tier Name**	**Tier Content**
Syllable	Syllable	STRING-UTF8
Token	Word	STRING-UTF8
Part-Of-Speech	POS	STRING-UTF8
Prosodic Units	PU	'PU'
Discourse Units	DU	{ 'DU', 'ADU'}

Creating prosodic units

French

- Both phonetic and phonological criteria have been used to segment
- 3 levels $\rightsquigarrow$ First evaluation $\rightsquigarrow$ Derive a less detailed but more reliable dataset
- Second Evaluation: κ-score of 0.71

Mandarin

- 1 level
- Cues: pitch reset (a shift upward in overall pitch level), lengthening, alternation of speech rate, occurrences of paralinguistic sounds
- Process
  - Train 3 labelers on 150 turns until a satisfactory consistency rate
  - Rest of the dataset was completed by the three labelers independently

Producing discourse units

- Discourse Segmentation guidelines inspired from [Muller et al., 2012] and [Chen, 2011]
- Combine
  - semantic criterion: main predicate (denoting an eventuality $\rightsquigarrow$ propositional content)
  - discourse criterion *(presence of discourse markers)*
  - pragmatic criterion *(recognition of specific speech acts)*
- Evaluation:
  - French: $0.74 < \kappa < 0.85$
  - Taiwan Mandarin: 0.86

Illustration

(1) French Discourse Units

[on y va avec des copains]$_{du}$ [on avait pris le ferry en Normandie]$_{du}$ [puisque j'avais un frère qui était en Normandie]$_{du}$ [on traverse]$_{du}$ [on avait passé une nuit épouvantable sur le ferry]$_{du}$

[we going there with friends]$_{du}$ [we took the ferry in Normandy]$_{du}$ [since I had a brother that was in Normandy]$_{du}$ [we cross]$_{du}$ [we spent a terrible night on the ferry]$_{du}$

(2) Mandarin discourse units

[qishi ta jiang de na ge ren yinwei ta you qu kai guo hui]$_{du}$ [ta hai you jiang]$_{du}$ [keneng shi ye bu zhidao wei she me]$_{du}$

[in fact the one he mentioned had the meeting]$_{du}$ [he said in addition]$_{du}$ [probably (he) did not know why, either]$_{du}$

Size of units

	dur (s)	# syll	#tokens	# PU
PU-fr	0.88	3.9	3.3	-
PU-tw	1.44	6.4	4.4	-
DU-fr	2.51	11.1	9.5	2.8
DU-tw	2.17	9.6	6.6	1.5

Table : Comparative size of the units produced

Association of prosodic and discourse units

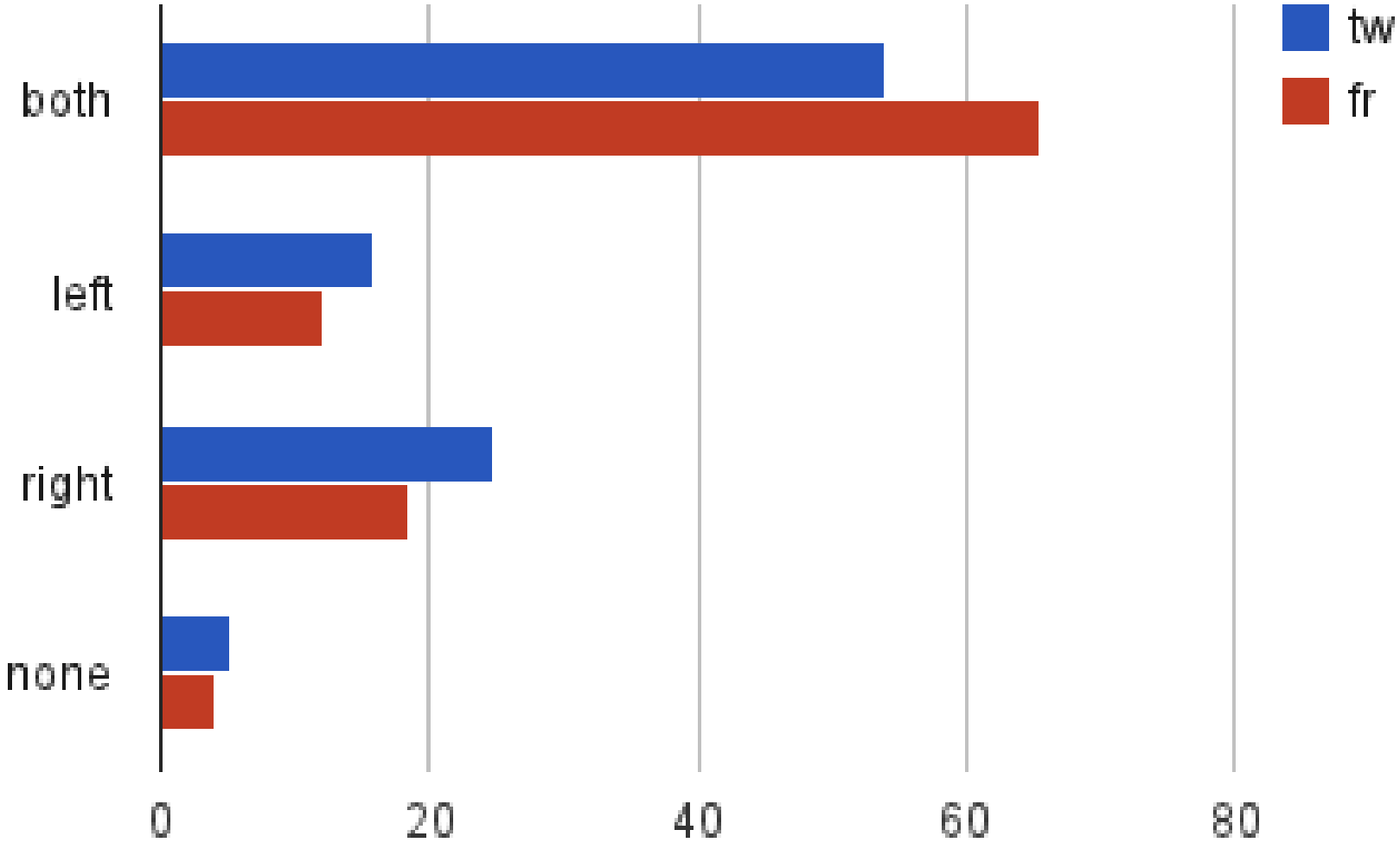


Figure : Distribution of PU/DU simplified association types

Syntactic categories at beginning boundaries

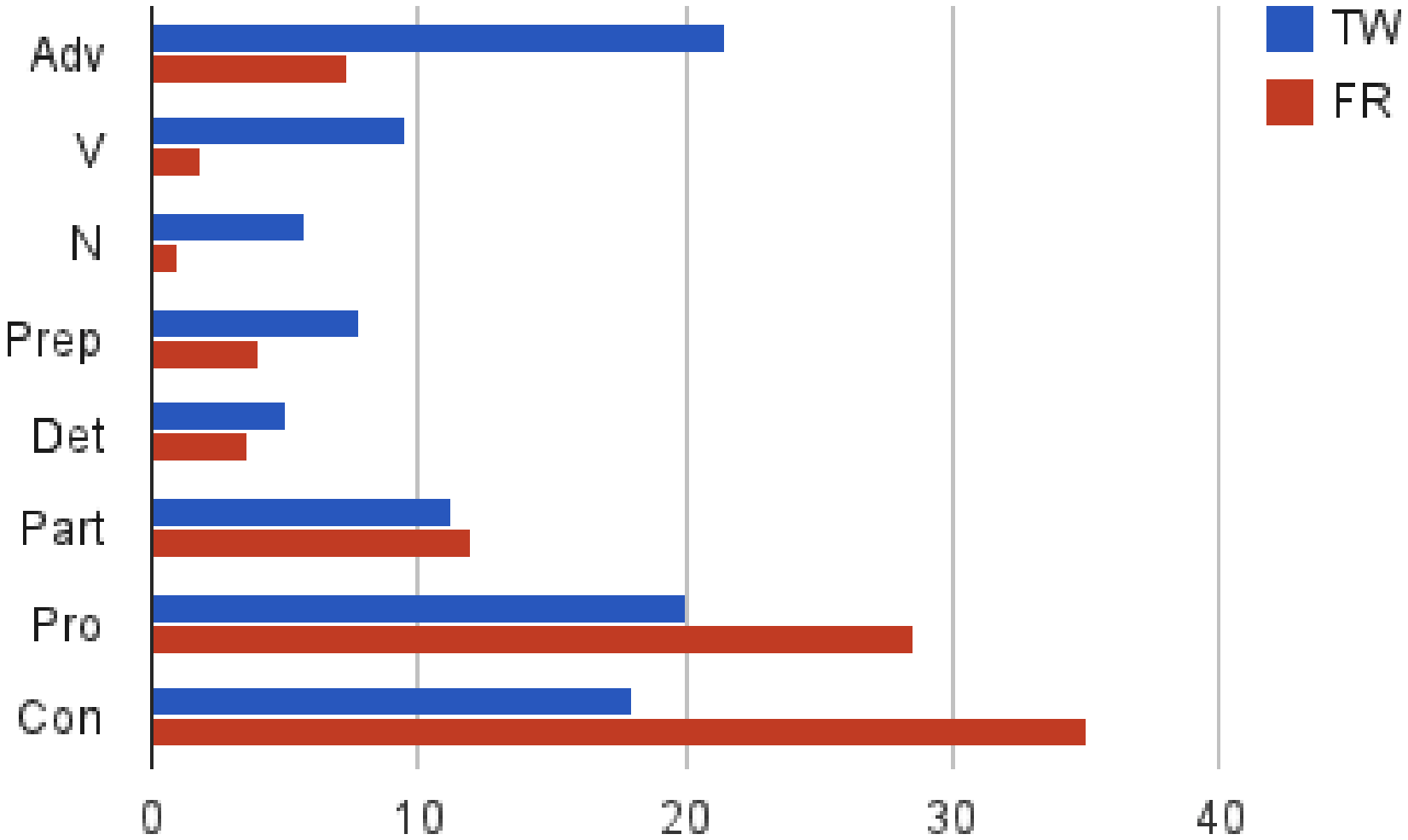


Figure : POS distribution at Initial matching boundaries

Syntactic categories at ending boundaries

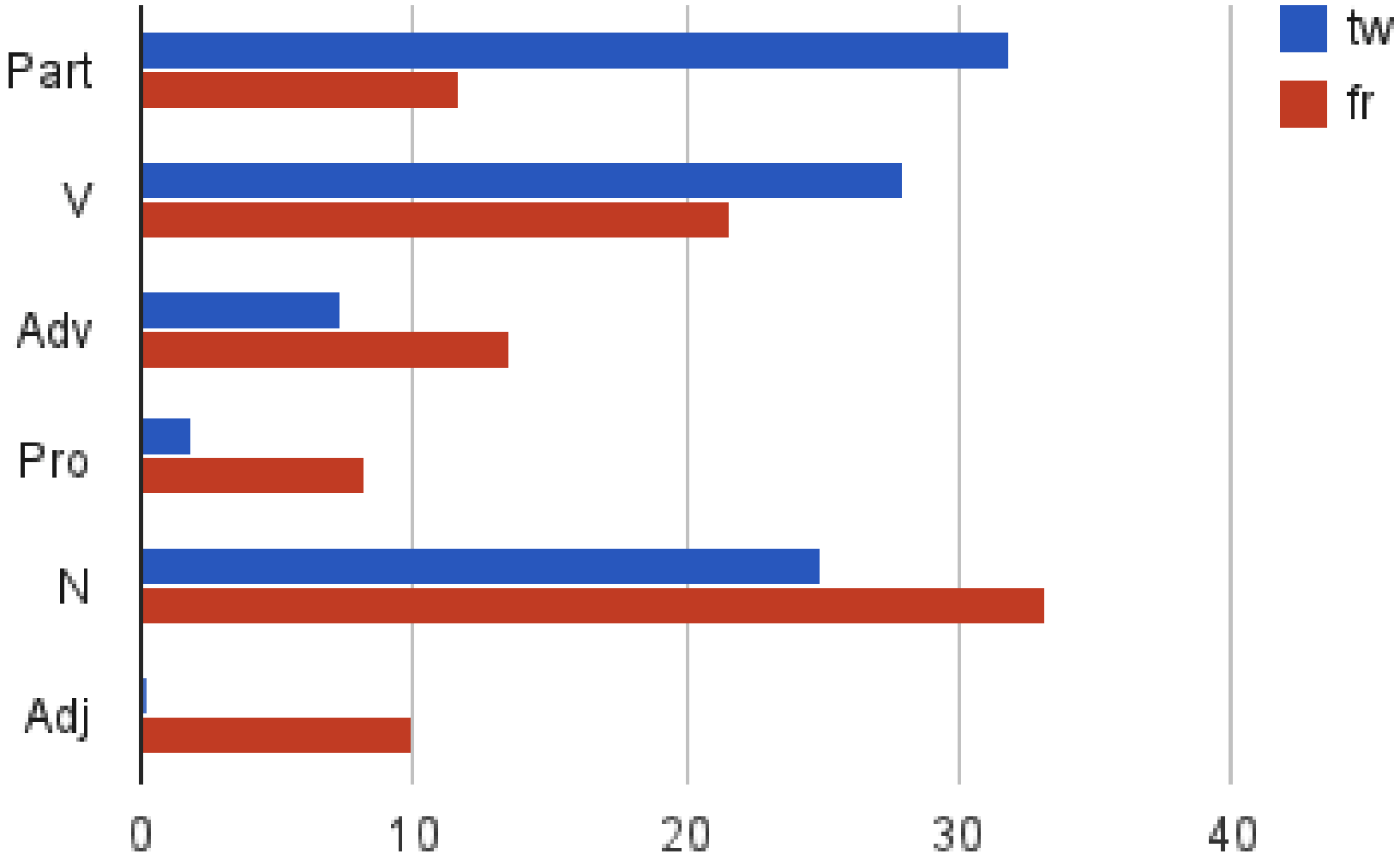


Figure : POS distribution at Final matching boundaries

Observations

Initial and starting 'tokens' fits more or less what is known a

- Mandarin
  - ∅-Anaphora extremely frequent in conversation
  - Initiale position = Topique (frequent construction)
  - Final particles are part of Mandarin grammar (aspect, mood,...)
- Français:
  - Initial Pronouns et Conjunctions (specially in conversation)

Chunks: a processing unit?

- Objective: define processing unit, "chunks" = first trial
- Hypothesis: *If chunks are processing units, the DUs and PUs across languages should remain similar in terms size-in-chunks distribution*
- Chunks: Created with hand-crafted rules based on POS tags
- Hypothesis not verified: different sizes across French and Taiwan Mandarin
  - Potential issue with sampling: turn-based selection vs. sequence-based selection
  - Comparability of the datasets?

Conclusion

- Very small differences in corpora design and annotation results in observable differences
- Comparable 'enough' dataset of significant size requires
  - ideally joint design + mutual checks at each corpus building decision point
  - achievable on a unique site only or thought deep and continuous collaboration

Ongoing / starting work:

- Systematic investigation Mono-,bi- and tri-chunks PUs and DUs
- Radical approach to QCIL

Radical approach to QCIL

- Non-supervised endogenous segmentation for both spoken french and mandarin (based on syllables)
- [Magistry and Sagot, 2012] approach and system
- 'spoken language' tagging, chunking and semantic analysis $\rightsquigarrow$ spoken structures
  - genre, putain : Discourse markers (not Nouns)
- cross-lingual mapping / comparison of spoken structures
  - made easier thanks to the radical approach sketched
  - through formal characterisations

Illustration of the first step

(3) et donc on s'installe un peu partout # on on allume les trucs
and so we settle down a bit everywhere # we we light up the things

a. [et donc on s'installe un peu partout] # [on on allume les trucs]
b. edo∼k o∼sU∼stAl U∼p@ pARtu o∼n o∼nAlym le tRyk

(4)

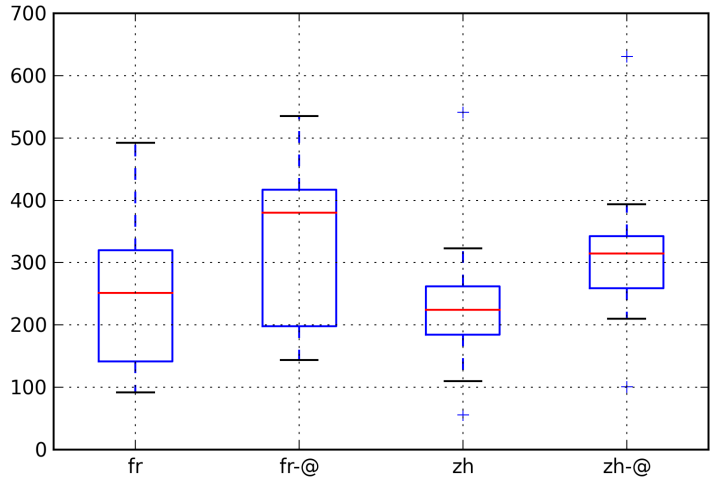
a. [edo∼k/DM o∼/Pro sU∼stAl/V U∼p@/R pARtu/R]
[o∼n/Pro o∼n/Pro AlymV/ le/Det tRyk/N]

(5) [edo∼k]$_{DC}$ [o∼ sU∼stAl]$_{VC}$ [U∼p@ pARtu]$_{RC}$ [o∼n o∼n Alym]$_{VC}$[le tRyk]$_{NC}$

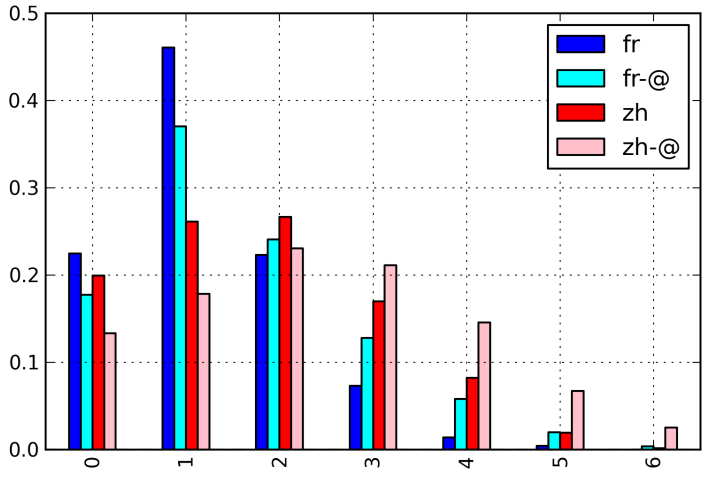
a. [edo∼k/*DC* VC RC] [VC NC]
b. [edo∼k/*DC* VC-action RC] [VC-action NC-generic]

The werewolf corpus

Comparative overview of a game

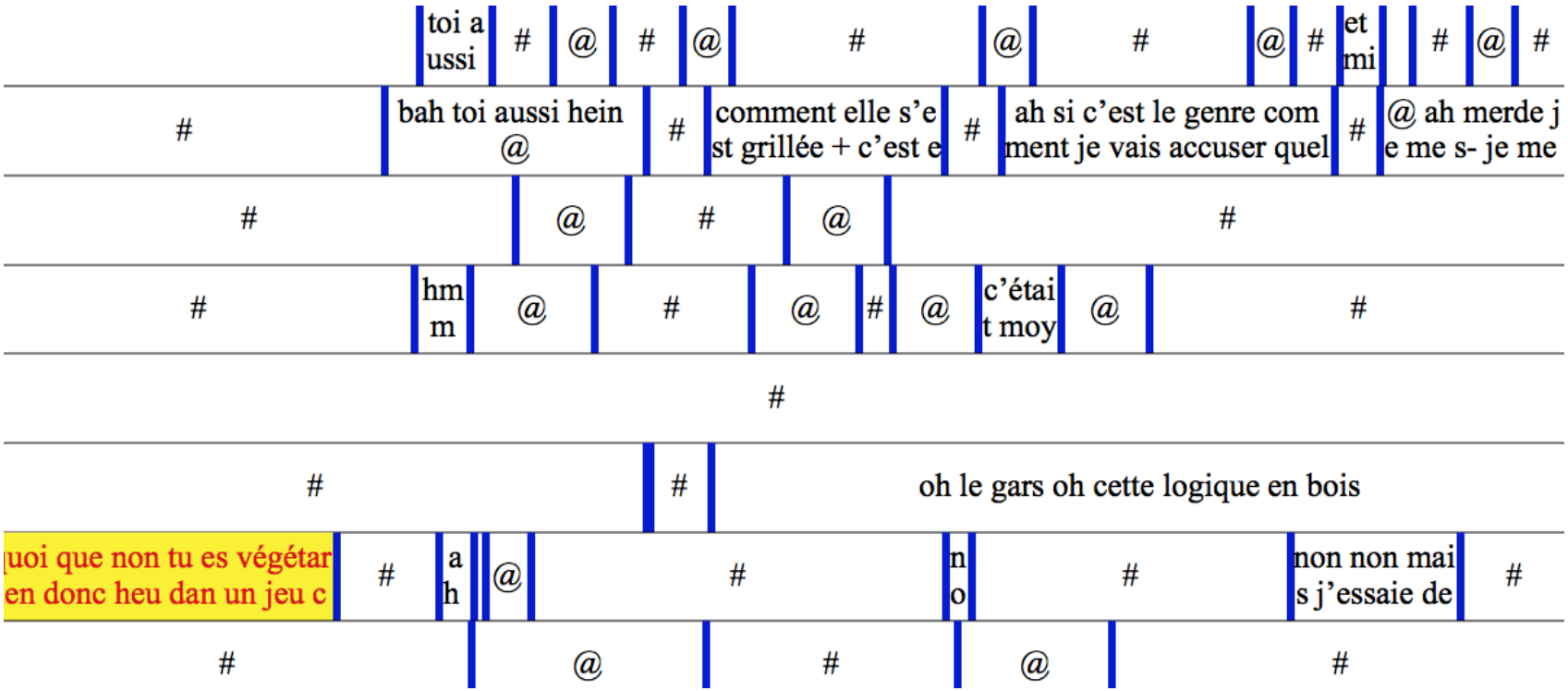


Actual Speaking Duration



# of simultaneous speakers

French illustration



Corpus interesting for

- Fiercely spontaneous and interactional language structures
- Perfectly comparable (when protocol will be fixed)
- Attitudes, Emotion (laughter)
- Deceptive speech, Argumentation
- Linguistic management of group evolution through the interaction

References I

Chen, A. C. (2011).

Prosodic phrasing in Mandarin conversational discourse: A computational-acoustic perspective.

PhD thesis, Graduate Institute of Linguistics, National Taiwan University.

Clancy, P. M., Thompson, S. A., Suzuki, R., and Tao, H. (1996).

The conversational use of reactive tokens in english, japanese, and mandarin.

Journal of pragmatics, 26(3):355–387.

Couper-Kuhlen, E. and Selting, M. (2001).

Introducing interactional linguistics.

Studies in interactional linguistics, 122.

Ginzburg, J. (2012).

The Interactive Stance: Meaning for Conversation.

Oxford University Press.

References II

Higuchi, S., Chaminade, T., Imamizu, H., and Kawato, M. (2009).
Shared neural correlates for language and tool use in broca's area.
Neuroreport, 20(15):1376–1381.

Lambrecht, K. (1988).
Presentational cleft constructions in spoken French.
Clause combining in grammar and discourse, pages 135–179.

Levinson, S. C. (2006).
On the human" interaction engine".
In *Wenner-Gren Foundation for Anthropological Research, Symposium 134*, pages 39–69. Berg.

Levitan, R., Benuš, Š., Gravano, A., and Hirschberg, J. (2015).
Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison.
In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 325.

References III

Magistry, P. and Sagot, B. (2012).

Unsupervized word segmentation: the case for Mandarin Chinese.

In *Proceedings of the 50th Annual Meeting of the ACL*, pages 383–387.

Maynard, S. K. (1990).

Conversation management in contrast: Listener response in Japanese and American English.

J. of Pragmatics, 14(3):397–412.

Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Farah, B., Bras, M., Draoulec, A. L., and Vieu, L. (2012).

Manuel d'annotation en relations de discours du projet annodis.

Technical Report 21, CLLE-ERS, Toulouse University.

Traugott, E. C. and Dasher, R. B. (2001).

Regularity in semantic change, volume 97.

Cambridge University Press.

References IV

Ward, N. and Tsukahara, W. (2000).

Prosodic features which cue back-channel responses in english and japanese.

Journal of pragmatics, 32(8):1177–1207.

Lexicon produced by the unsupervised segmenter for our French corpus

- si tu veux / ça doit / je crois / tu vois / tu sais
- et puis / non mais / enfin bon / ah ouais
- une fois / des fois
- pour faire
- en même temps
- comme si