Coreference resolution

Amalia Todirascu
University of Strasbourg
todiras@unistra.fr

Coreference resolution

- Coreference resolution => aims to identify all the expressions referring the same discourse entity
- Difficult task
  - Complex linguistic knowledge (morphologic and syntactic properties)
  - External/domain-specific knowledge
- Used for multiple NLP applications : text simplification, dialogue systems, text summarization…

Definitions

- **Coreference relation** : 2 expressions are coreferent if they are instances of the same entity
- **Coreference chains** : at least 3 referring expressions identifying the same entity (Schnedecker, 1997)
- **Singleton** : one referring expression
- **Referring expressions (mentions)** :
  - proper names : Justin Pierre James Trudeau
  - noun phrases : a Canadian politician (indefinite NP), the prime minister of Canada, the leader of the Liberal Party (definite NP), this prime minister (demonstrative NP)
  - Pronouns : he (personal pronouns), who (relative pronouns), himself (demonstrative pronouns)
  - Other : demonstrative determiner (his)

  - « Justin Pierre James Trudeau ([...]) is a Canadian politician who has served as the prime minister of Canada since 2015, the 23rd since Confederation, and has been the leader of the Liberal Party since 2013. » (Wikipedia, Friday 9th April 2021)

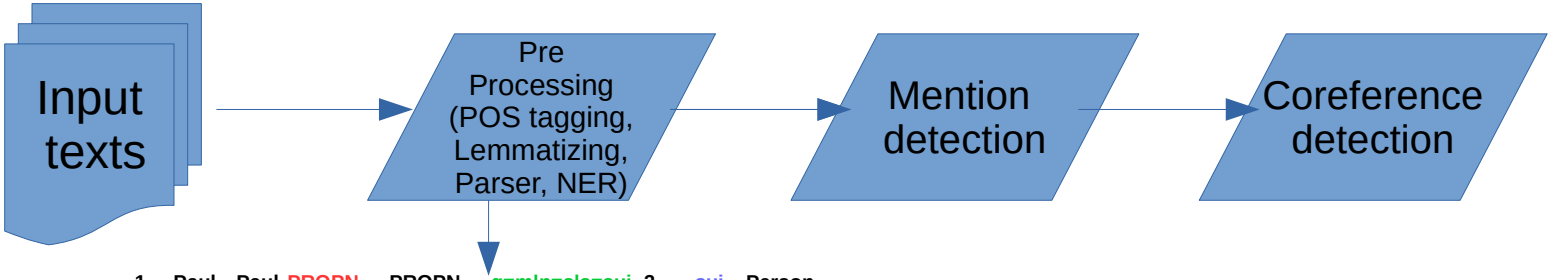
A Difficult Task (I)

- “Le prince Philip n'est plus. Comme annoncé dans un communiqué rédigé par Buckingham Palace, le mari de la reine Elisabeth II s'est éteint dans la matinée de ce vendredi 9 avril. Âgé de 99 ans, il serait parti en paix après des derniers mois plutôt difficiles. En effet au début du mois de mars il avait été hospitalisé pour une intervention en raison d'un problème cardiaque. Une disparition qui émeut tout le Royaume-Uni, profondément attaché à cette figure de la monarchie. A commencer par son petit-fils William et son épouse Kate Middleton, qui lui ont rendu hommage sur leurs réseaux sociaux. De leur côté Meghan et Harry ne se sont pour le moment pas encore exprimés, probablement en raison du décalage horaire. Tous deux se sont installés à Los Angeles, en Californie, depuis qu'ils ont quitté la famille royale.” (Voici.fr, ven. 9 avril 2021 à 2:47 PM, read 9th April 2021)

A Difficult Task (II)

- Problems
  - Several discourse entities :
    - Single individuals : *Prince Philip, son petit-fils William, son épouse Kate Middelton, Elisabeth II, Harry, Meghan*
    - Groups : *Harry et Meghan, son petit-fils William et son épouse Kate Middelton, la famille royale*
  - Various expressions designing the same entity : pronouns, Nps, proper names …
  - External knowledge :
    - {le mari de la reine Elisabeth II, prince Philip}
    - {famille royale => monarchie}
  - Morphological constraints :
    - Le prince Philip (ms) => il (ms)
    - le prince Philip (ms) => cette figure de la monarchie (fs)
  - Events : « Prince Philip n'est plus » …. « cette disparition »

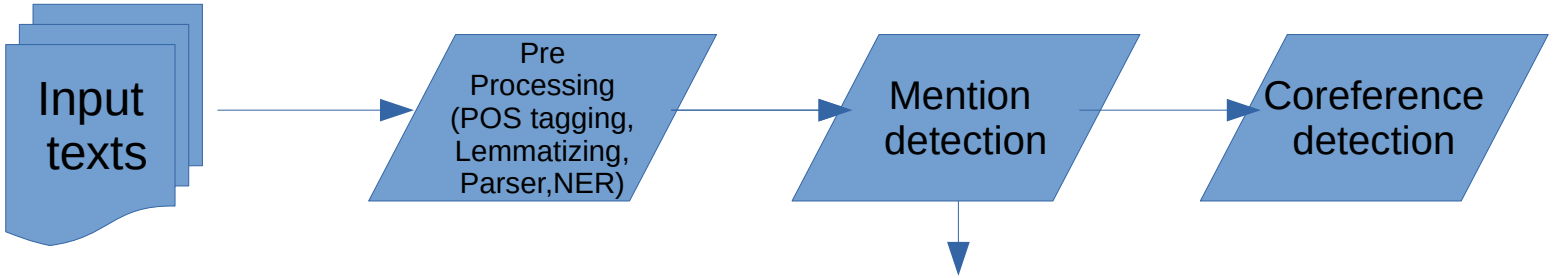
A Coreference Resolution System



1	Paul	Paul	PROPN	PROPN	g=m|n=s|s=suj	3	suj	Person	_	
2	ne	ne	ADV	ADV	s=neg	3	mod	_	_	
3	laisse	laisser	V	V	m=ind|n=s|p=3|t=pst	0	root	_	_	
4	à	à	P	P	_	3	a_obj	_	_	
5	personne	personne	NC	NC	g=m|n=s|p=3|s=c	4	obj.p	_	_	
6	le	le	DET	DET	g=m|n=s|s=def	7	det	_	_	
7	soin	soin	NC	NC	g=m|n=s|s=c	3	obj	_	_	
8	de	de	P	P	_	7	dep	_	_	
9	sculpter	sculpter	VINF	VINF	m=inf	8	obj.p	_	_	
10	sa	son	DET	DET	g=f|n=s|p=3|s=poss	11	det	_	_	
11	statue	statue	NC	NC	g=f|n=s|s=c	9	obj	_	_	
12	pour	pour	P	P	_	9	mod	_	_	
13	l'	la	DET	DET	g=f|n=s|s=def	14	det	_	_	
14	histoire	histoire	NC	NC	g=f|n=s|s=c	12	obj.p	_	_	

(Mind The Gap, Coavoux et Crabbé, 2017)

A Coreference Resolution System

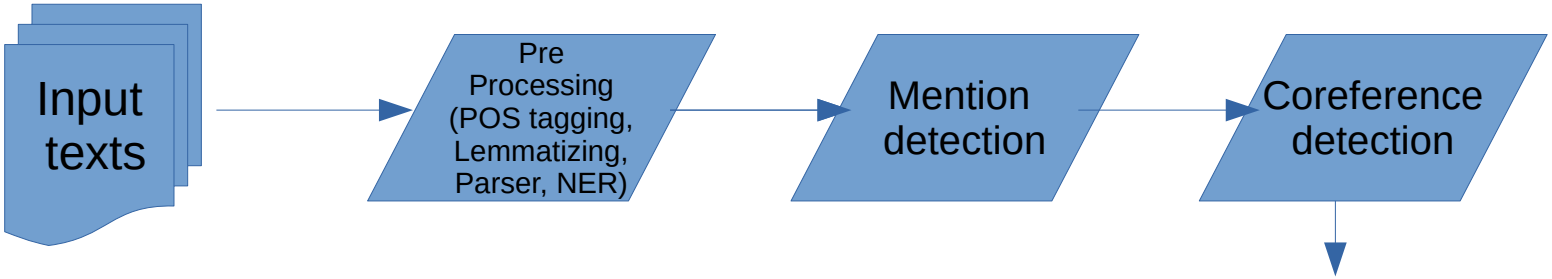



```
1   Paul       Paul      PROPN  PROPN  g=m|n=s|s=suj  3   suj    (1)
2   ne         ne        ADV    ADV    s=neg          3   mod
3   laisse     laisser   V      V      m=ind|n=s|p=3|t=pst 0  root
4   à          à         P      P      _              3   a_obj
5   personne   personne  NC     NC     g=m|n=s|p=3|s=c    4   obj.p
6   le         le        DET    DET    g=m|n=s|s=def  7   det    (2
7   soin       soin      NC     NC     g=m|n=s|s=c    3   obj
8   de         de        P      P      _              7   dep
9   sculpter   sculpter  VINF   VINF   m=inf          8   obj.p  _    2)
10  sa         son       DET    DET    g=f|n=s|p=3|s=poss 11  det   (3|(5)
11  statue     statue    NC     NC     g=f|n=s|s=c    9   obj    3)
12  pour       pour      P      P      _              9   mod
13  l'         la        DET    DET    g=f|n=s|s=def  14  det    (4
14  histoire   histoire  NC     NC     g=f|n=s|s=c    12  obj.p  _    4)
```

Detected mentions

Paul
Le soin de sculpter sa
sa statue
l'histoire

A Coreference Resolution System



Coreference chains

e1= {Paul, sa}
e2= {le soin de sculpter}
e3={sa statue}
e4= {l'histoire}

```
1   Paul       Paul       PROPN  PROPN  g=m|n=s|s=suj  3   suj    (1)
2   ne         ne         ADV    ADV    s=neg          3   mod
3   laisse     laisser    V      V      m=ind|n=s|p=3|t=pst 0  root
4   à          à          P      P      _              3   a_obj
5   personne   personne    NC     NC     g=m|n=s|p=3|s=c    4   obj.p
6   le         le         DET    DET    g=m|n=s|s=def  7   det    (2
7   soin       soin       NC     NC     g=m|n=s|s=c    3   obj
8   de         de         P      P      _              7   dep
9   sculpter   sculpter   VINF   VINF   m=inf          8   obj.p  _   2)
10  sa         son        DET    DET    g=f|n=s|p=3|s=poss 11  det    (3|(1)
11  statue     statue     NC     NC     g=f|n=s|s=c    9   obj    3)
12  pour       pour       P      P      _              9   mod
13  l'         la         DET    DET    g=f|n=s|s=def  14  det    (4
14  histoire   histoire   NC     NC     g=f|n=s|s=c    12  obj.p  _   4)
```

Coreference Resolution

- symbolic approaches (rule-based approaches)
  - pronoun/anaphora resolution (Hobbs, 1978 ; Mitkov, 2002 ; Lappin & Leass, 1994)
  - coreference resolution : constraint checking rules (Oberle, 2019)
- Statistical machine learning approaches
  - Supervised (Ng, 2010) vs non-supervised approaches (Haghighi & Klein, 2009)
  - Coreference annotated corpora : OntoNotes (Hovy et al, 2006), ACE (Walker et al, 2006)
- Deep learning approaches (Lee et al, 2019 ; Kantor et Globerson, 2019)
  - Word embeddings (BERT (Devlin et al 2018))
  - Coreference annotated corpora

Rule-based Approaches

- Algorithm
  - Identification of new entities (Proper names, indefinite NPs)
  - Selection of anaphor mentions (pronouns, demonstrative NPs, some definite NPs)
  - Pair creation
  - Check linguistic constraints between possible antecedents and anaphors
  - Selecting pairs which high evaluation score : the type of referring expressions (Accessibility Theory :Ariel, 1990) + various constraints

Rule-based Approaches (II)

- Linguistic constraints
  - (Mitkov, 2002) (Lapin and Leass, 1994):
    - same number and gender
    - short distance between the pronoun and its antecedent
    - type of referring expressions
- Discourse constraints :
  - Similar syntactic function
  - Frequency
  - Salience
- Semantic constrains
  - Centering theory : Grosz et al, 1995 ; Accessibility theory : Ariel, 1990; Constraint theory : (Beaver, 1990)

New Entities

- “{Le prince Philip} n'est plus. Comme annoncé dans {un communiqué rédigé par {Buckingham Palace}}, le mari de la {reine Elisabeth II} s'est éteint dans la matinée de ce vendredi 9 avril. Âgé de 99 ans, il serait parti en paix après des derniers mois plutôt difficiles. En effet au début du mois de mars il avait été hospitalisé pour une intervention en raison d'un problème cardiaque. Une disparition qui émeut tout {le Royaume-Uni}, profondément attaché à cette figure de la monarchie. A commencer par {{son petit-fils William} et {son épouse Kate Middleton}}, qui lui ont rendu hommage sur leurs réseaux sociaux. De leur côté {{Meghan} et {Harry}} ne se sont pour le moment pas encore exprimés, probablement en raison du décalage horaire. Tous deux se sont installés à {Los Angeles}, en {Californie}, depuis qu'ils ont quitté {la famille royale}.” (Voici.fr, ven. 9 avril 2021 à 2:47 PM, read 9th April 2021)

Proper nouns, indefinite NPs, some definite NPs : introducing new entities :low accessibility (Ariel, 1990)

Coreference Detection

- “{Le prince Philip} n'est plus. Comme annoncé dans {un communiqué rédigé par {Buckingham Palace}}, {le mari de la {reine Elisabeth II}} {s'}est éteint dans {la matinée de {ce vendredi 9 avril}}. Âgé de 99 ans, {il} serait parti en paix après {des derniers mois plutôt difficiles}. En effet {au début du mois de mars} {il} avait été hospitalisé pour {une intervention en raison d'un problème cardiaque}. {Une disparition} {qui} émeut tout {le Royaume-Uni}, profondément attaché à {cette figure de la monarchie}. A commencer par {{son petit-fils William} et {son épouse Kate Middleton}}, qui lui ont rendu hommage sur leurs réseaux sociaux. De leur côté {{Meghan} et {Harry}} ne se sont pour le moment pas encore exprimés, probablement en raison du décalage horaire. Tous deux se sont installés à {Los Angeles}, en {Californie}, depuis qu'ils ont quitté {la famille royale}.” (Voici.fr, ven. 9 avril 2021 à 2:47 PM, read 9th April 2021)

Definite NPs, pronouns, demonstrative NPs : highly accessible entities (Ariel, 1990)

{le mari de la {reine Elisabeth II}} => {Le prince Philip}, {un communiqué…} {Buckingham Palace}
Hum+, m, s, subject Hum+, m, s, subject Abs+, m, s, obl Place+, m, s, obj

Rule-based Approaches (III)

- ODACR (Oberle, 2019) following (Longo, 2013)
  - Complex tokenisation : NER, specific MWE
  - Complex linguistic information : POS tagging, lemma, dependency parsing : Talismane (Urieli, 2013)
  - Named entities : YAGO
  - Rule detecting hyponyms/hyperonyms links: Mon chat … cet animal
  - Rules detecting groups : Paul et Marie, ...ils, … le couple

Rule-based Approaches

- Advantages
  - Useful for low resourced languages
  - Linguistic interpretation
- Drawbacks
  - Hand made rules
  - Time-consuming engineering
  - incoherences

ML : Supervised Approaches

- Several models (Ng, 2010)
  - Mention-pairs models (Soon et al. 2001 ; Ng and Cardie 2002)
    - Binary classification : coreferent pairs or non-coreferent pairs
    - Transitive relations : (m1,m2) and (m1,m3) are coreferent => (m1,m2,m3) represent the same entity
    - Vector of features : number, genre, syntactic function, lexical head, type of referring expressions, same surface elements, named entities, semantic relations between the heads
  - Mention ranking : RECONCILE (Stoyanov et al, 2010)
  - Entity-mention : link a mention to one single entity (similarity distances)
- Large corpora with multiple annotations : morphosyntactic, syntactic, semantic, NE, coreference
  - English : OntoNotes (Hovy et al, 2006), ACE (Walker et al, 2005)
  - French : ANCOR (oral transcriptions) (Muzerelle et al, 2015), DEMOCRAT (written texts) (Landragin, 2016)

ML : Supervised Approaches (II)

- Features
  - Same number : *le chat (ms) – il (ms)*
  - Same gender : *Elisabeth II (fs) – elle (fs)*
  - Same syntactic function
  - Same lexical head : *le chat – ce chat*
  - Strict lexical matching : *Kate Middelton – Kate Middelton*
  - Partial lexical matching : *ce vilain animal – l'animal*
  - Distance
  - Speaker
  - Frequency
  - salience
  - Semantic features : hyperonymy/hyponymy relations *(le chat - l'animal*)

Mention-pair Model

- “{Le prince Philip} n'est plus. Comme annoncé dans {un communiqué rédigé par {Buckingham Palace}}, {le mari de la {reine Elisabeth II}} {s'}est éteint dans {la matinée de {ce vendredi 9 avril}}. Âgé de 99 ans, {il} serait parti en paix après {des derniers mois plutôt difficiles}. En effet {au début du mois de mars} {il} avait été hospitalisé pour {une intervention en raison d'un problème cardiaque}. {Une disparition} {qui} émeut tout {le Royaume-Uni}, profondément attaché à {cette figure de {la monarchie}}. A commencer par {{son petit-fils William} et {son épouse Kate Middleton}}, {qui} {lui} ont rendu hommage sur {leurs} réseaux sociaux. De leur côté {{Meghan} et {Harry}} ne se sont pour le moment pas encore exprimés, probablement en raison du décalage horaire. Tous deux se sont installés à {Los Angeles}, en {Californie}, depuis qu'ils ont quitté {la famille royale}.” (Voici.fr, ven. 9 avril 2021 à 2:47 PM, read 9th April 2021)

{le mari de la {reine Elisabeth II}}, {Le prince Philip} => coreferent
Hum+, m, s, subject Hum+, m, s, subject
{un communiqué…} {Buckingham Palace} => non-coreferent
{Le prince Philip}, {il} => coreferent
Hum+, m, s, subject m,s, 3rd, subject

ML : Supervised Approaches (II)

- Hybrid method : selecting mentions by constraints (Lee et al, 2011)
  - English, Chinese, Arabic : https://nlp.stanford.edu/software/dcoref.shtml
- Binary classification (Desoyer et al, 2015)
  - Mention pair model, SVM algorithm
  - 2 classes : coreferent mentions and non-coreferent mentions
  - Large set of features : type of the referring expression, lexical head, strict matching, number, gender, speaker…
  - Assume that the mention detection is already done

ML Approaches

- Advantages
  - Less linguistic engineering rather then symbolic approaches
  - Linguistic interpretation
- Drawbacks
  - Large annotated corpora
  - Simplified annotation model

Deep-learning Approaches

- Mention detection : splitting the texts into a set of spans (Lee et al, 2019) (Kantor and Globerson, 2019)
- Coreference resolution : using word embeddings and character embeddings
  - BERT (Devin et al, 2018)
  - Elmo

cofr

- An end-to-end coreference resolution system for French
  - https://github.com/boberle/cofr
  - Models for mention detection and for coreference resolution
  - In the framework of DEMOCRAT and ALECTOR projects
- Neural approach :
  - An adaptation from (Kantor and Globerson, 2019)
  - French corpora : DEMOCRAT (Landragin, 2016), ANCOR (Muzerelle et al)

Annotated Corpus : DEMOCRAT

- DEMOCRAT project : https://www.ortolang.fr/market/corpora/democrat
- Large corpus
  - from 12th to 20th century
  - Narrative (litterature, newspapers) vs informative texts
  - Various genres : novels, newspaper articles, technical reports, encyclopedia articles
  - Several domains : law, biology, finance
  - Text fragments (10000 words)
- Method : at least 2 annotators per text
  - Manual annotation tools : SACR (Oberle, 2018) : https://github.com/boberle/sacr, TXM (URS plugin :http://textometrie.ens-lyon.fr/?lang=en), Analec (Landragin, 2014)
  - Specific guidelines (mention annotation and coreference annotation)
  - Mention detection (including singletons)
  - Relating the mentions into coreference chains

Annotated Corpus with SACR

[#3] FrancoisHollande François Hollande , à Paris Paris 5Avril le 5 avril . Photo AudoinDesforges Audoin Desforges pour Liberation Libération

[#4] Jovial et inentamé, FrancoisHollande l'ex-président publie 10Avril2018 mercred LivreBilan un livre bilan , «LeconsDuPouvoir les Leçons du pouvoir ». FrancoisHollande Il revendique Quinquennat FrancoisHollande son quinquennat , FrancoisHollande concède QuelquesErreurs quelques erreurs et FrancoisHollande déplore ExhibitionDeSaViePrivee l'exhibition de ViePrivee FrancoisHollande sa vie privée .

Evaluation

- MUC Score (Vilainet al., 1995)
  - Link based: Counts the number of common links and computes f-measure
- CEAF (Luo2005); entity based
- BLANC (Recasens and Hovy 2011): counts coreference links and non-coreference links
- All of them use precision, recall and f-mesure (harmonic average)
  - Precision (P): % of elements in a hypothesized reference chain that are in the true reference chain
  - Recall (R): % of elements in a true reference chain that are in the hypothesized reference chain
- State-of -the art systems: ~85% (MUC)

References

Corblin, F. (1995). Les formes de reprise dans le discours. Anaphores et chaînes de référence. Presses universitaires de Rennes.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Désoyer, A., Landragin, F., & Tellier, I. (2015). Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français: le système CROC.

Grobol, L. (2019). Neural coreference resolution with limited lexical context and explicit mention detection for oral French.

Grosz, B. J., Joshi, A., and Weinstein, S. (1995) Centering: a framework for modelling the local coherence of discourse. Computational Linguistics, 21(2), pp. 44–50. MIT Press.

Hobbs, J. R. (1982). Towards an understanding of coherence in discourse. Strategies for natural language processing, 223-244.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: the 90% solution. In Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers (pp. 57-60).

Kantor, B., & Globerson, A. (2019). Coreference resolution with entity equalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 673-677).

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. Computational linguistics, 20(4), 535-561.

References (II)

- Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT).
- Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045.
- Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. arXiv preprint arXiv:1804.05392.
- Longo, L. (2013). Vers des moteurs de recherche" intelligents": un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence (Doctoral dissertation)
- Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J. Y., Pelletier, A., Maurel, D., & Villaneau, J. (2014, May). ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures.
- Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 1396-1411).
- Oberle, B. (2019) Détection automatique de chaînes de coréférence pour le français écrit : règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques. Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL) 2019, Jul 2019, Toulouse, France.
- Schnedecker, C., & Landragin, F. (2012). Les chaînes de référence: présentation. Langages, (3), 3-22.
- Stoyanov, V., Babbar, U., Gupta, P., & Cardie, C. (2011). Reconciling OntoNotes: Unrestricted Coreference Resolution in OntoNotes with Reconcile. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (pp. 122-126).
- Walker, Christopher, et al. ACE 2005 Multilingual Training Corpus LDC2006T06. Web Download. Philadelphia: Linguistic Data Consortium, 2006.
- Wilkens, R., Oberle, B., Landragin, F., & Todirascu, A. (2020). French coreference for spoken and written language. In Language Resources and Evaluation Conference (LREC 2020) (pp. 80-89).