

Non Deterministic Chunking

François Trouilleux

LRL, Université Blaise-Pascal, Clermont-Ferrand

NooJ Conference, June 2009, Tozeur

Outline

- Definitions and state of the art
- Problems chunking an ambiguous input
- Towards a new application mode for NooJ grammars ?

Work based on the development of a chunker for French,
to be presented in TALN 2009

What is a chunk ?

- Abney's definition (1991):

I define chunks in terms of *major heads*. Major heads are all **content words** except those that appear between a **function word** *f* and the **content word** that *f* selects. For example, *proud* is a major head in [**a** *man*] [**proud**] [**of his son**], but *proud* is not a major head in [**the proud man**], because it appears between the function word *the* and the content word *man* selected by *the*.

A French example

<GP>Pour cent francs</GP> <GP>par an</GP>,
<NV>elle faisait</NV> <GN>la cuisine</GN>
et <GN>le ménage</GN>, <NV>cousait</NV>,
<NV>lavait</NV>, <NV>repassait</NV>, <NV>savait</NV>
<NV>brider</NV> <GN>un cheval</GN>, <NV>engraisser</NV>
<GN>les volailles</GN>, <NV>battre</NV> <GN>le beurre</GN>,
et <NV>resta</NV> <GA>fidèle</GA>
<GP>à sa maîtresse</GP>...

- Span almost the whole text
- No embedding
- Most frequent types: *function** *content*

(GP, GN, etc., are the categories of the EASY evaluation campaign)

Chunking vs morphology

- Different notations for the same information

<i>isolated words</i>		<i>affix-like spelling</i>
give it to me	donne- le - moi	dá me lo (es)
a sea	une mer	mare (ro)
the sea	la mer	mare a
the blue sea	la mer bleue	albastra a mare
in the house	dans la maison	a ház ban (hu)

In French, in principle, function words on the right-hand side are connected with an hyphen

State-of-the-art chunkers

- *Rule-based* chunkers: (Hindle, 1994), (Kinyon, 2001), (Aït-Mokhtar *et al.*, 2002), (Bourigault *et al.*, 2005), etc.
- Deterministic
 - **either:** <GN>**la petite brise**</GN> <NV>**la glace**</NV>
(en: *the little breeze freezes her*)
 - **or:** <GN>**la petite**</GN> <NV>**brise**</NV> <GN>**la glace**</GN>
(en: *the little (one) breaks the ice*)
- Incremental, two-step process
 1. Part-of-speech disambiguation
la,DET **petite**,A **brise**,N **la**,CL **glace**,V
 2. Pattern matching
GN = DET, A*, N ;
NV = CL*, V ;

Limitations

- Determinism is inadequate:
 - real ambiguities exist
 - the information to deal with some spurious ambiguities is not always available
- Incremental process:
 - errors in the first stage affect the second stage
 - the linguistic description is redundant

- Definitions and state of the art
- **Problems chunking an ambiguous input**
- Towards a new application mode for NooJ grammars ?

Goal

- Produce all possible analysis in terms of chunks using essentially information on the internal composition of chunks (100% recall)

la,DET la,CL la,N	petite,ADJ petite,N	brise,N brise,V	la,DET la,CL la,N	glace,N glace,V
GN			NV	
GN		NV	GN	
		GN		

The diagram illustrates the goal of producing all possible analyses of a sentence in terms of chunks. The sentence is "la petite brise la glace". The table shows the possible chunkings and their internal compositions. Green arrows indicate dependencies between chunks across rows.

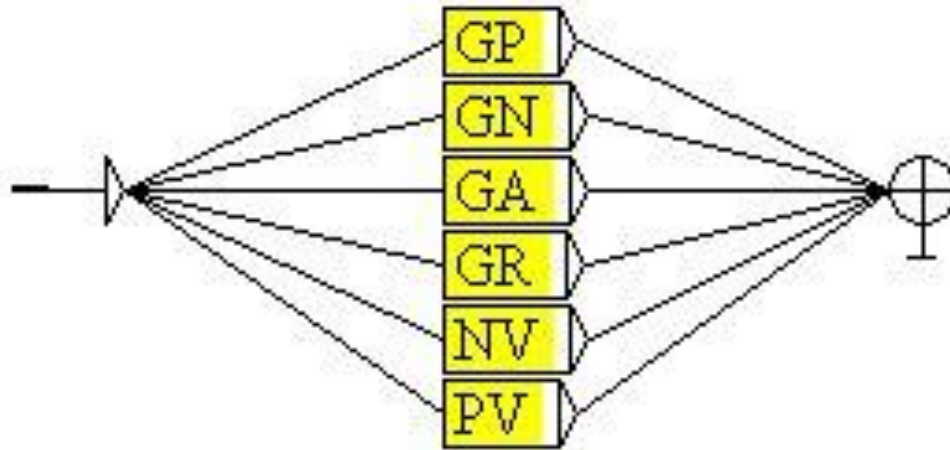
Goal

- Produce all possible analysis in terms of chunks using essentially information on the internal composition of chunks (100% recall)
- **With reasonable precision**

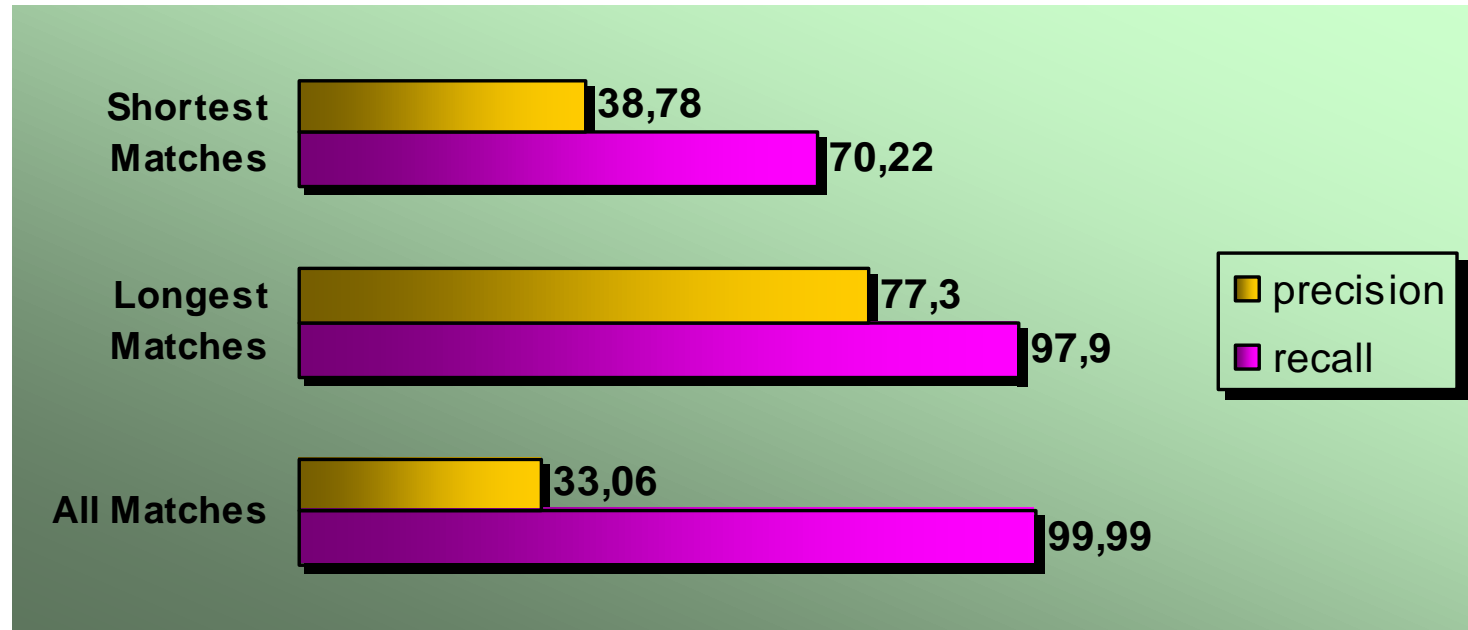
la,DET la,CL la,N	petite,ADJ petite,N	brise,N brise,V	la,DET la,CL la,N	glace,N glace,V
GN			NV	
GN		NV GN	GN	
GN	GA GN GN		GN	NV GN

The basic model

- Assume a grammar which specifies the various types of chunks
- Suppose it is correct wrt to chunk composition
- Apply it on corpus using the different modes



Results of the basic model



- \approx 22,500 word corpus, 11,144 chunks
- 22,561 precision errors in the *All Matches* mode
- 234 recall errors in the *Longest Matches* mode

All Matches Problems

il,CL	la,CL la,DET la,N	porte,N porte,V
NV		

All Matches Problems

il,CL	la,CL la,DET la,N	porte,N porte,V
NV		
	NV	
		NV

All Matches Problems

il,CL	la,CL la,DET la,N	porte,N porte,V
NV		
	NV	
		NV
	GN	

All Matches Problems

il,CL	la,CL la,DET la,N	porte,N porte,V
NV		
	NV	
		NV
	GN	
	GN	GN

100% recall (1/1)
14% precision (1/7)

Longest Match Problems

- 203/233 errors (87%) due to an extension of the right frontier of GN/GP one lexical word too far:
 - <GN>**les deux premiers la**</GN> <NV>**frôlaient**</NV>
 - <GN>**la classe ouvrière**</GN>
 - <GN>**des bas gris**</GN>
 - <GN>**le surnaturel est**</GN> **tout simple**
- 28 errors (12%) with pattern: tout|toute, <WF>
 - <GN>**tout change**</GN>, <GN>**toute surprise**</GN>
- 2 due to a punctuation problem
 - En voilà <GN>**une Mme Lehoussais**</GN>, qui au lieu de prendre un jeune homme...

- Definitions and state of the art
- Problems chunking an ambiguous input
- Towards a new application mode for NooJ grammars ?

New grammar application mode

- Given a grammar, let LF be the set of (declaratively specified) left-hand side function words, e.g.

LF = <PREP>+<CL>+<DET>+<NUM>+<NEG>;

- Chunking mode:
for each text unit, select the sets of segments which **maximize** both the **number of initial LFs** and **input coverage**

(initial LFs = LFs before the head of the chunk)

Example 1

il,CL	la,CL la,DET la,N	porte,N porte,V
NV		
	NV	
		NV
	GN	
	GN	GN

#initial LF	cov.
2	3
1	2
0	1
1	2
0	2

Example 2

les,DET les,CL	deux,NUM	premiers,ADJ	la,DET la,CL la,N	frôlaient,V

Example 2

les,DET les,CL	deux,NUM	premiers,ADJ	la,DET la,CL la,N	frôlaient,V
GN	1	GA	0	
GN			2	
GN				2

Example 2

les,DET les,CL	deux,NUM	premiers,ADJ	la,DET la,CL la,N	frôlaient,V
GN	1	GA	0	
GN			2	
GN			2	

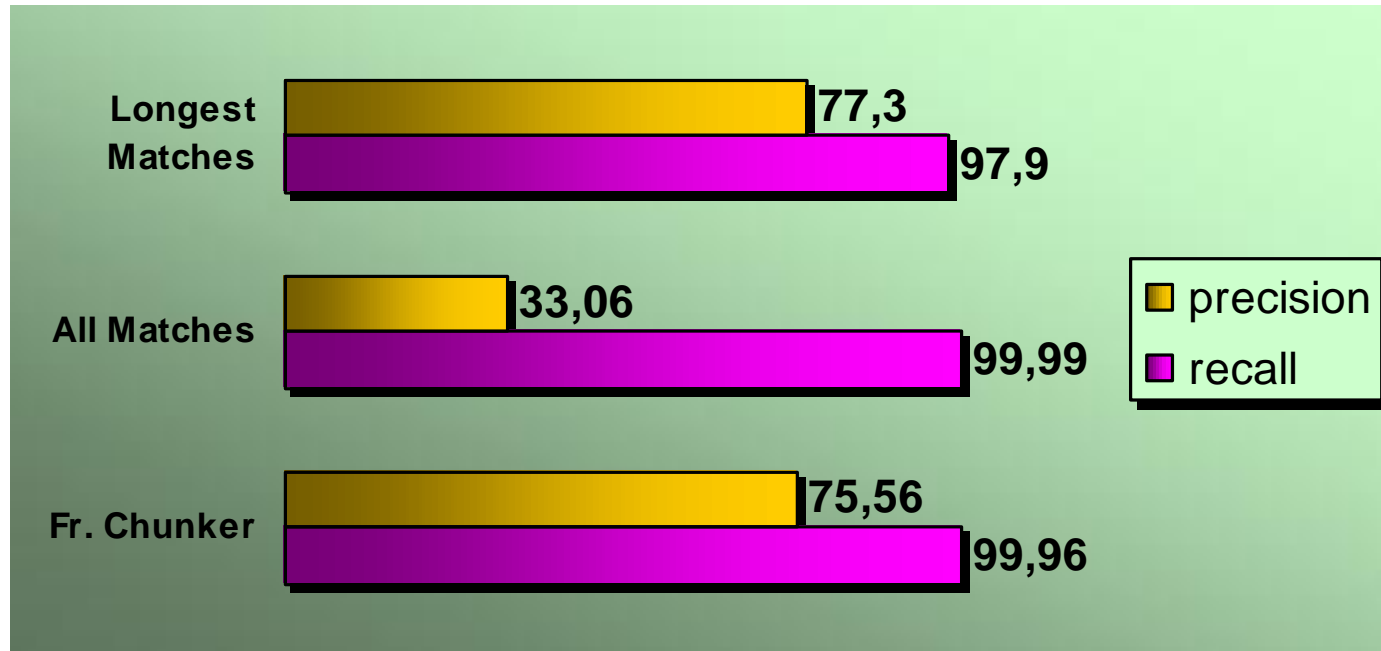
Example 2

les,DET les,CL	deux,NUM	premiers,ADJ	la,DET la,CL la,N	frôlaient,V
GN	1	GA	0	
GN			2	NV 1
GN			GN 0	NV 0
GN			2	NV 0

Example 2

les,DET les,CL	deux,NUM	premiers,ADJ	la,DET la,CL la,N	frôlaient,V
GN	1	GA	0	
GN			2	NV 1
			GN 0	NV 0
GN			2	NV 0

Expected result



- One may venture that the proposed mode would give results similar to my French chunker.
- Possibly interesting application time

Questions

- Would it be difficult to implement ?
- What about other languages ?
- ...