

AC/DC project:

Acesso a Corpora/Disponibilização de Corpora (Corpus Access and Availability)

- 1st phase: all publicly available corpora Web served in a uniform way
- 2nd phase: parsing those corpora with a robust parser and Web serve them
- 3rd phase: process the corpora with further tools/parsers

Computational Processing of Portuguese

A government funded initiative to foster R&D in the area

Aims at creating a Distributed Resource Center for Portuguese

- Resource creation or improvement
- Resource cataloguing
- Resource evaluation

<http://www.portugues.mct.pt>

AC/DC: Differences in token interpretation

Processing stage	EBRANOT with and without punctuation		NATPANOT with and without punctuation	
	Original version	891,280	664,194	860,373
Parser's output	908,003	670,797	897,770	730,418
MWE translation	924,010	686,804	913,037	745,721
Contractions merge	880,466	643,916	849,981	683,117
MWE expansion	897,116	660,566	864,293	697,429
Clitics merge	890,987	654,440	862,560	695,698

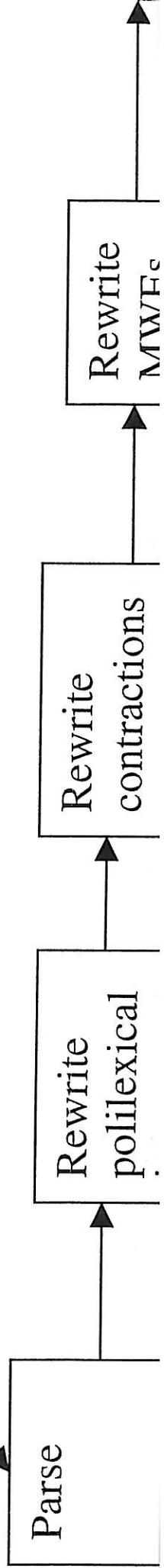
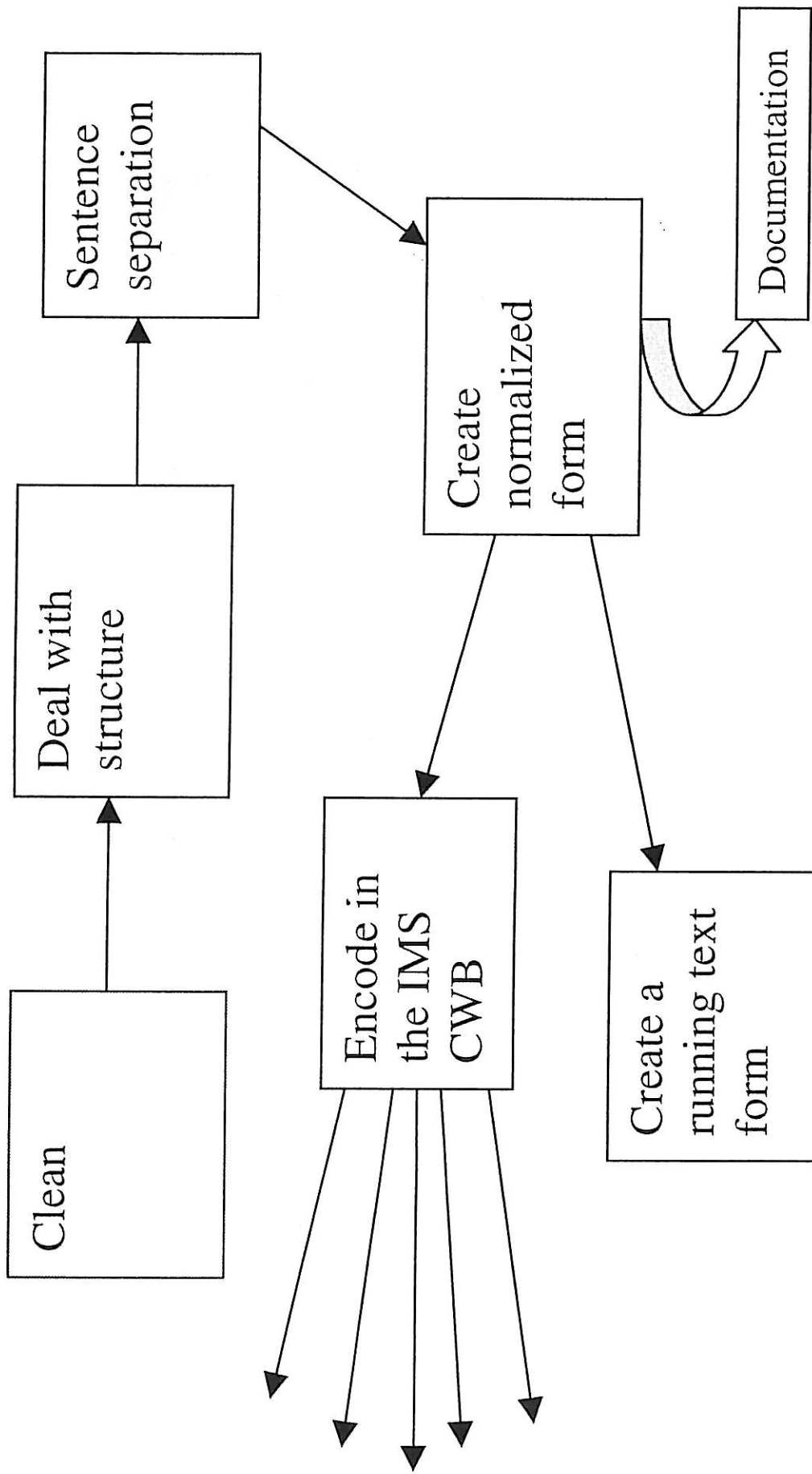
Tokenization is to a large extent arbitrary – or theory dependent. What is a word? (what is a sentence?) How to deal with proper nouns?

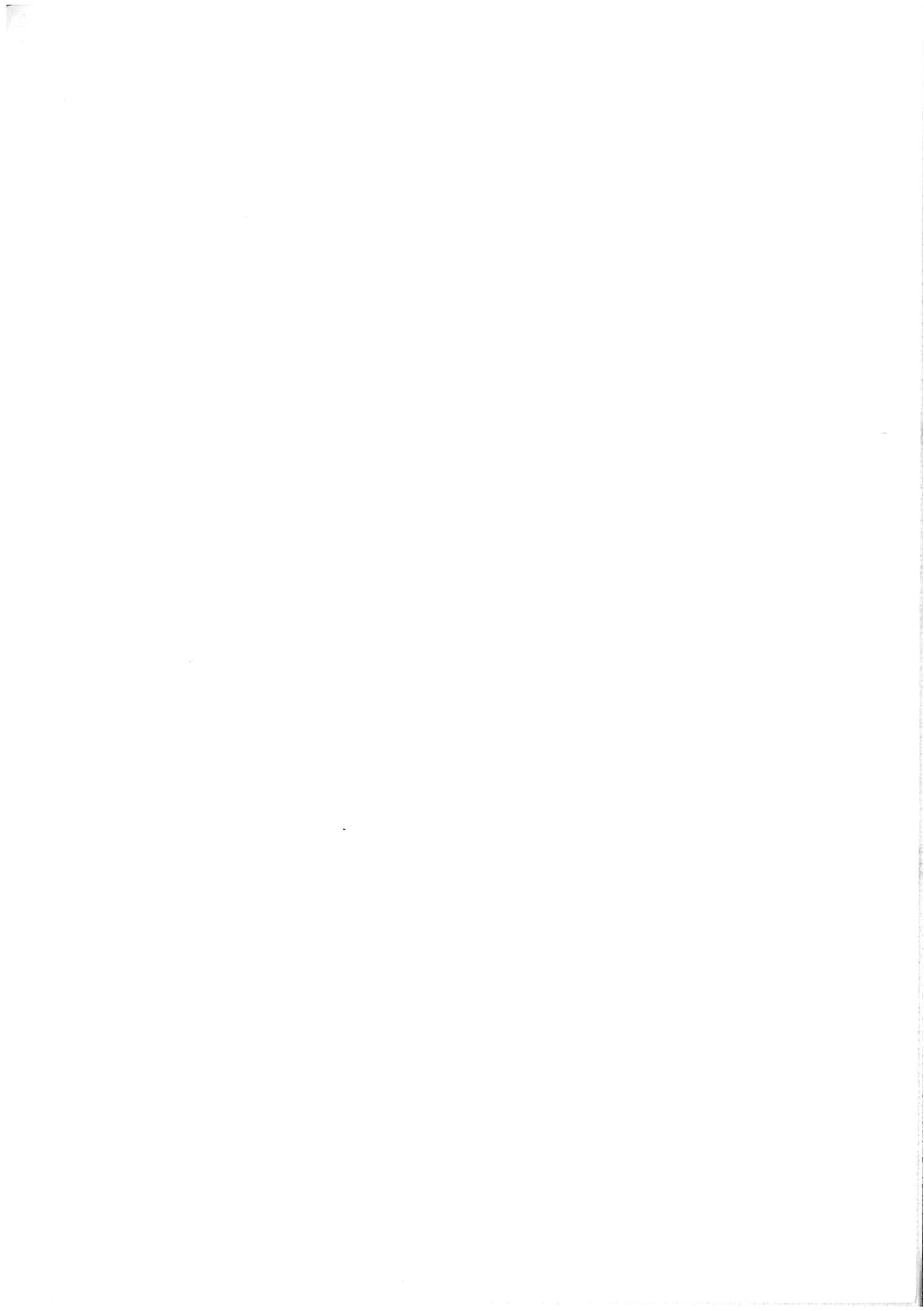
AC/DC project: Parsed corpora

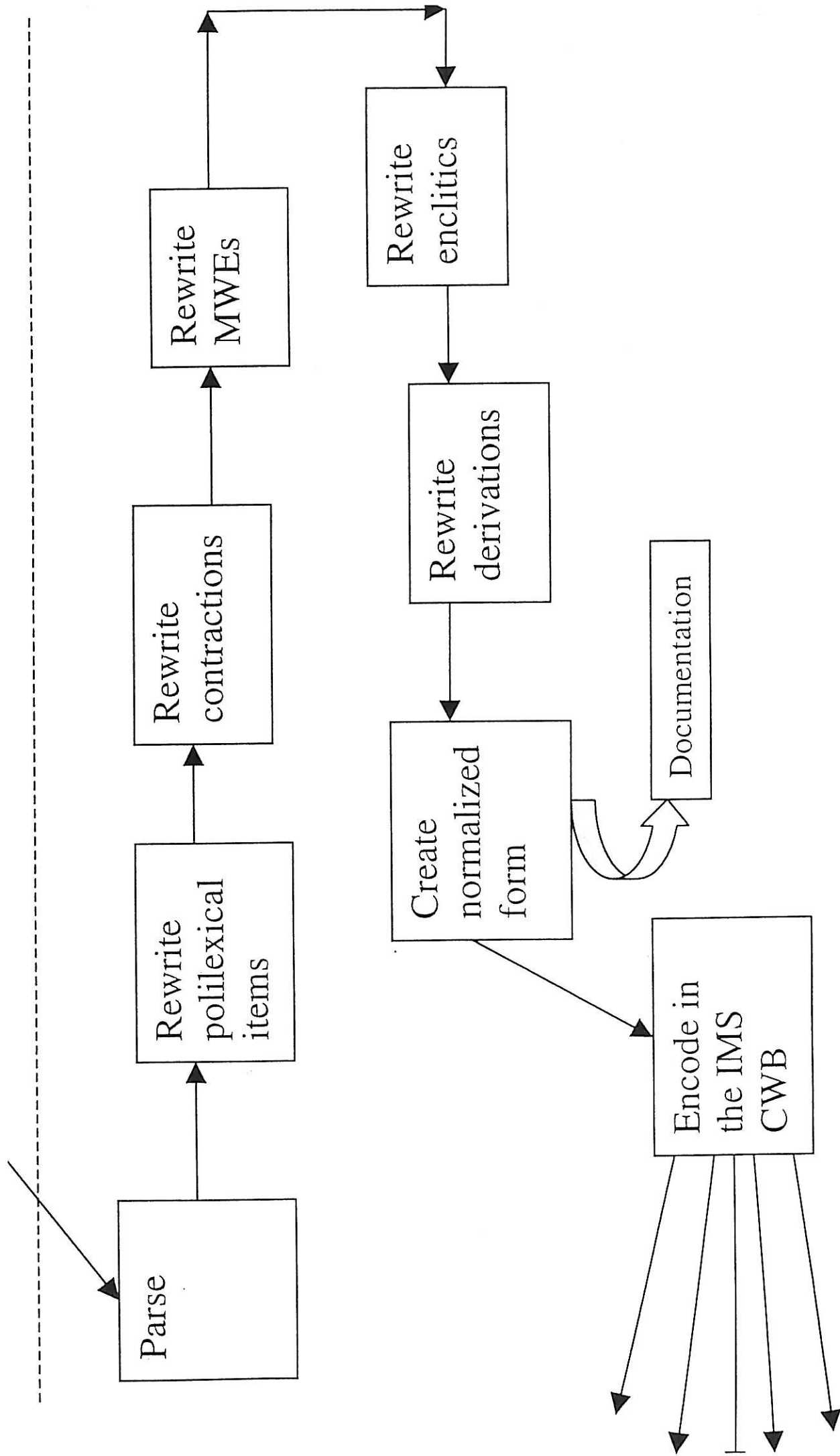
As of 25th May 2000:

(NATPANOT: v. 2.0, 24 May 2000; EBRANOT: v. 2.0, 24 May 2000)

Parsed corpus	NATPANOT	%	EBRANOT	%
Sentences	226,323		45,480	
Words	6,279,741		722,743	
((Common) Nouns	1,336,618	21.8	146,302	20.2
Verbs	813,148	12.9	118,110	16.3
Adjectives	375,789	6.0	42,086	5.8
Adverbs	312,013	5.0	48,068	6.7
Proper nouns	550,273	8.8	32,373	4.5
Contractions	497,255	7.9	43,377	6.0







- note the AC/DC genie widely known

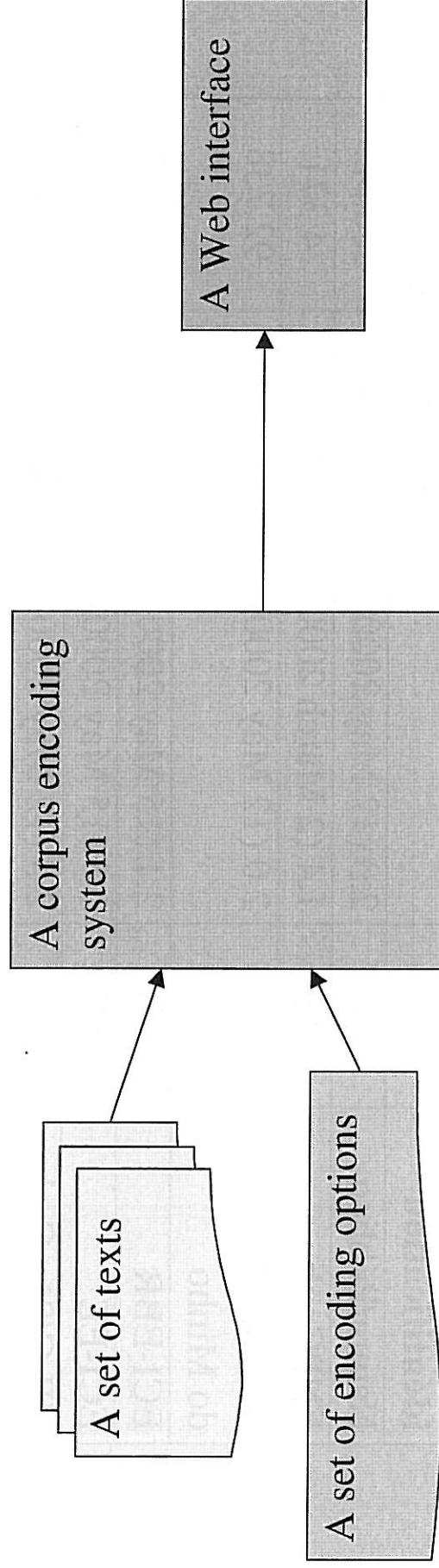
- report the on-going cooperation between the Comp. Proc. of Port project represented by myself as the IRL project

~~the interesting~~ rather some interesting items

Corpora

AC/DC interface

IMS Corpus Workbench



AC/DC project: Corpora involved

As of 25th May 2000:

Corpus identification	Variant	Version	Size in words (k) ~	Size in sentences
Natura/Público	PP	2.0 (12 May 2000)	6,255	226,323
ENPCpub	M	1.4 (2 March 2000)	72	4,371
Natura/Diário do Minho	PP	2.0 (12 May 2000)	2,115	92,738
ECI-EBR	PB	2.1 (15 May 2000)	721	45,613
ECI-EE	PP	1.5 (15 May 2000)	26	780
NILC/São Carlos	PB	2.5 (12 May 2000)	33,789	2,226,514
FrasesPP	PP	1.4 (15 May 2000)	16	594
FrasesPB	PB	1.3 (15 May 2000)	19	652
Total			43,013	2,597,585

AC/DC project: Shortcomings

- Lack of documentation
- Lack of a real user community: how many people do corpus-based analysis of Portuguese in the world?
- Some of the corpora have more problems than normal text (capitalization failures in long sections; mixture of very different genres; encoding errors; duplicated items)

This is the reason for two longer-term goals:

- creating larger resources, using the AC/DC framework, but also catering for physical distribution on CD
- creating Web-based learning materials on Portuguese corpus studies

