Corpuscle :: Documentation of KIAP (Kulturell Identitet i Akademisk Prosa)

KIAP Corpus attributes

Attribute name	Case-insensitive search by default	Scope	Description	Comment
word	yes	cpos	Token - word, punctuation or tag	
type		cpos	Type of token – **w** (word), **p** (punctuation) or **t** (tag)	
filename		main	Original filename	See more information in the **Extratextual elements (metadata)** table below
author		main	Author of the article	See more information in the **Extratextual elements (metadata)** table below
authorship		main	Author gender - **masc**, **fem** or **several**	See more information in the **Extratextual elements (metadata)** table below
source		main	The source of the scientific article	See more information in the **Extratextual elements (metadata)** table

				below
language		main	The language of the scientific article - **english**, **french** or **norwegian**	See more information in the **Extratextual elements (metadata)** table below
discipline		main	The discipline of the scientific article - **economics**, **linguistics** or **medicine**	See more information in the **Extratextual elements (metadata)** table below
article		main	Tells if the article is in IMRAD format - **imrad** or **non-imrad**	See more information in the **Extratextual elements (metadata)** table below
path		cpos	The XML path of the token	

Tags

Main element

Tag	Type	Tag attributes	Description
<main>	Non-empty		Contains all the elements for one article

Extratextual elements (metadata)

Tag	Type	Tag attributes	Description
<meta>	Non-empty		Contains the following elements with metadata about the article
<filename>	Non-empty		Contains the alphanumerical label given to the article. Also original filename, less extension. Same as the **filename** corpus attribute
<author>	Non-empty		Contains the name(s) of the author(s). Same as the **author** corpus attribute
<authorship>	Non-empty		Specifies if the article was written by one woman (**fem**), one man (**masc**) or by several persons (**several**). Same as the **authorship** corpus attribute
<source>	Non-empty		Contains information about the journal that the article was taken from (name, volume, issue, pages). Same as the **source** corpus attribute
<language>	Non-empty		Specifies the language that the article was written in - **english**, **french** or **norwegian**. Same as the **language** corpus attribute
<discipline>	Non-empty		Specifies the article discipline - **linguistics**, **economics** or **medicine**. Same as the **discipline** corpus attribute

<article>	Non-empty		Specifies IMRAD vs. non-IMRAD articles (values: **imrad** vs **non-imrad**). Same as the **article** corpus attribute

Article structure

Tag	Type	Tag attributes	Description
<title>	Non-empty	@type	Contains the/a title of the article. Some French and Norwegian articles have an English title in addition to their native title. The @type xml attribute has a value of **native** or **foreign** accordingly
<abstract>	Non-empty	@type	Contains an abstract for the article. Abstracts may be given in the native language or, if it is French or Norwegian, in English - or both or none of them. (A few French articles also have an abstract in Dutch.) The @type xml attribute has a value of **native** or **foreign** accordingly. The abstract is sometimes placed after the <body> of the article
<body>	Non-empty		The main structural element of the article, containing the article itself. Body-matter structural tags are <body>, <intro>, <mmr>, <disc>, <mid> and <concl>. All articles have a <body> part
<intro>	Non-empty		If an introductory text part is structurally identifiable on the basis of headings or other layout features, it is contained in an <intro> element. Otherwise, the beginning of the text is treated as part of the mid section of the article.

<mid>	Non-empty		Medical articles are typically structured according to the IMRAD format, which means that there are sections dealing with material (and methods) and results followed by a discussion. The encoding recognises this by dividing the middle into a part containing materials, methods and results, contained in an <mmr> element, and a discussion an a <disc> element. In some articles, no such division has been feasible, and the whole mid part has been included in either the <mmr> or the <disc> section depending on which label is the more appropriate. Economics and linguistics articles are more variably structured, and a <mid> tag pair replaces the two tag pairs for medical articles, i.e. <mmr> and <disc>. Thus, article bodies from these disciplines are maximally divided into <intro>, <mid> and <concl>. Not all articles have any identifiable concluding section. This is more often the case for medicine, but articles from other disciplines, too, may lack a conclusion. Therefore, the body may end with the middle part
<mmr>	Non-empty		See <mid>

<disc>|Non-empty||See <mid>|

<concl>	Non-empty		See <mid>
<notes>	Non-empty	@type	Contains notes. Footnotes have also been placed here. The @type xml attribute tells if the note is an endnote (**end**) or a footnote (**page**, sometimes erroneously encoded as **foot**)

<references>	Non-empty		Contains bibliography or references. All articles have a <references> element
<misc>	Non-empty		Contains front and back matter that can not be classified as belonging to more specific elements - appendices, acknowledgements, contact addresses, epigraphs, etc.

Textual elements

Tag	Type	Tag attributes	Description
<subtitle>	Non-empty		Contains a section heading. The elements <subtitle>, <quote>, <example> and <table> contain portions that are more or less outside the text proper, or, in the case of <quote>, are not the author's own words
<quote>	Non-empty		Contains a direct quote of at least three words; the reference to the quoted text is not included. See also <subtitle>
<example>	Non-empty		Contains, in linguistic articles, linguistic examples that have been put in separate paragraphs or consist of more than one word . See also <subtitle>
<table>	Non-empty		Contains tables, figures, mathematical formulae and similar matter; figure captions and similar texts belonging to such features. The aforemented are thus excluded from the body text. See also <subtitle>

Other encoding

The code **nRRR** is used in front of numbers for bibliographic references in medical articles (where **n** is the variable number of works referred to) and **NNN** in front of note numbers.

Design & implementation: Paul Meurer, Uni Computing, 2014