| lvl | container | tag | opt | description | type | source |
|----------------------|------------------------|---|-----|---|----------------------|-----------------------|
| 10 | document | document | | container for document | dict | TET |
| | | | | created at (DD.MM.YY | | |
| 11 | document | createdAt | | HH:MM:SS) | str | system |
| 11 | document | createdBy | | created by module | str | system |
| | | - | | last modified at (DD.MM.YY | | , |
| 11 | document | modifiedAt | Х | HH:MM:SS) | str | system |
| | document | modifiedBy | Х | last modified by module | str | system |
| - | document | noFonts | X | number fonts | int | TET |
| - | document | noFooters | | number footers per page | int | LT |
| | | noHeaders | | number headers per page | int | LT |
| | document | noLines | | number lines in document | int | TET |
| | document | noLinesFooter | | number footer lines | int | LT |
| | document | noLinesHeader | | number header lines | int | LT |
| | document | noLinesHeading | Х | number heading lines | int | LT |
| - | | noLinesListBullet | X | number bulleted list lines | int | LT |
| | | noLinesListNumber | X | number numbered list lines | int | LT |
| | document | noLinesTable | ^ | number table lines | int | LT |
| | document | noLinesToc | Х | number toc lines | int | LT |
| | document | noListsBullet | X | number toc intes | int | LT |
| - | document | noListsNumber | X | number numbered lists | int | LT |
| | document | | | number pages in document | int | TET |
| | document | noPages | | number pages in document | IIIC | IEI |
| 11 | document | noParas | | number paragraphs in document | int | TET |
| | document | nor ar ar | | manusci paragraphis in accument | | |
| 11 | document | noSentences | Х | number sentences in document | int | spaCy |
| | document | noTables | | number tables in document | int | LT |
| | document | noWords | | number words in document | int | TET |
| | document | config | Х | container for configuration | dict | TET |
| | document | fonts | X | container for fonts | list | TET |
| | document | headings | ^ | container for headings | list | TET |
| | document | listsBullet | | container for headings | list | TET |
| | document | listsNumber | | container for numeredd lists | list | TET |
| | document | | | | list | TET |
| | document | pages params | v | container for pages container for parameters | dict | TET |
| | config | config | X | container for configuration | dict | TET |
| | Coming | Coming | X | container for configuration | uict | TT |
| | config | narcer | | container for module parcer | dic+ | 111 |
| | config | parser | | container for module parser | dict | |
| 21 | - | | | | | |
| 21 | config | tokenizer | | container for module tokenizer | dict | TET |
| | config | tokenizer | | container for module tokenizer include the configuration data in | dict | TET |
| | - | | | container for module tokenizer include the configuration data in the JSON file | | |
| 22 | config | tokenizer jsonInclConfig | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON | dict | TET |
| 22 | config | tokenizer | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON file | dict | TET |
| 22 | config config config | tokenizer jsonInclConfig jsonInclFonts | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON file include the heading data in the | dict bool bool | TET setup setup |
| 22 | config | tokenizer jsonInclConfig | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON file include the heading data in the JSON file | dict | TET |
| 22 22 22 | config config config | <pre>jsonInclConfig jsonInclFonts jsonInclHeading</pre> | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON file include the heading data in the JSON file include the bulleted list data in | dict bool bool | setup setup setup |
| 22 22 22 | config config config | tokenizer jsonInclConfig jsonInclFonts | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON file include the heading data in the JSON file include the bulleted list data in the JSON file | dict bool bool | TET setup setup |
| 22 22 22 22 | config config config | <pre>jsonInclConfig jsonInclFonts jsonInclHeading</pre> | | container for module tokenizer include the configuration data in the JSON file include the font data in the JSON file include the heading data in the JSON file include the bulleted list data in | dict bool bool | setup setup setup |

Version: 0.9.8 Page: 1/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|--------------------------------|-----|--|------|--------|
| | | | | include the parameters in the | | |
| 22 | config | jsonInclParams | | JSON file | bool | setup |
| | | | | include the sentences in the | | |
| 22 | config | jsonInclSentences | | JSON file | bool | setup |
| | | | | Improves the readability of the | | |
| 22 | config | jsonIndent | | JSON file | int | setup |
| | , | | | | | |
| 22 | config | jsonSortKeys | | Sort the keys in ascending order | bool | setup |
| | 3 | 5.000 | | maximum Levenshtein distance | | |
| 22 | config | ltFooterMaxDistance | | for a footer line | int | setup |
| | config | ltFooterMaxLines | | maximum number of footers | int | setup |
| | 8 | TOT GO GETTIAND THE B | | Maximum Levenshtein distance | | setup |
| 22 | config | ltHeaderMaxDistance | | for a header line | int | setup |
| | COTTING | TelleadelMaxDistalice | | Tor a ricader line | 1110 | эстар |
| 22 | config | ltHeaderMaxLines | | Maximum number of headers | int | setup |
| | comig | ItheadermaxLines | | number of lines following the | IIIC | setup |
| | | | | | | |
| 22 | config | l tuo adima Eila In al Na Ctor | | heading to be included as context into the JSON file | haal | cotup |
| | config | ltHeadingFileInclNoCtx | | | bool | setup |
| | | | | number of lines following the | | |
| | | ltHeadingFileInclRegex | | heading to be included as | | |
| | config | p | | context into the JSON file | bool | setup |
| | <u>.</u> | | | maximum level of the heading | | |
| 22 | config | ltHeadingMaxLevel | | structure | int | setup |
| | | | | minimum number of pages to | | |
| 22 | config | ltHeadingMinPages | | determine the headings | int | setup |
| | | | | file with rules to determine the | | |
| 22 | config | ltHeadingRuleFile | | headings | bool | setup |
| | _ | | | tolerance of vertical indentation | | |
| 22 | config | ltHeadingToleranceLlx | | in percent | int | setup |
| | _ | | | minimum number of entries to | | |
| 22 | config | ltListBulletMinEntries | | determine a bulleted list | int | setup |
| | _ | | | file with rules to determine the | | |
| 22 | config | ltListBulletRuleFile | | bulleted lists | bool | setup |
| | | ltListBulletToleranceL | | tolerance of vertical indentation | | |
| 22 | config | lx | | in percent | bool | setup |
| | | | | | | |
| | | | | if it is set to true, the regular | | |
| | | ltListNumberFileInclRe | | expression for the numbered list | | |
| 22 | config | gexp | | is included in the JSON file | bool | setup |
| | | | | | | |
| | | | | minimum number of entries to | | |
| 22 | config | ltListNumberMinEntries | | determine a numbered list | bool | setup |
| | | | | file with rules to determine the | | |
| 22 | config | ltListNumberRuleFile | | numbered lists | bool | setup |
| | | ltListNumberToleranceL | | tolerance of vertical indentation | | |
| 22 | config | lx | | in percent | bool | setup |

Version: 0.9.8 Page: 2/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-------------|-------------------------------------|-----|--------------------------------------|----------|--------|
| | | | | | | |
| | | | | if it is set to true, the empty ells | | |
| | | ltTableFileInclEmptyCo | | are included in the separate | | |
| 22 | config | lumns | | JSON file with the tables | bool | setup |
| | | | | maximum number of pages for | | |
| | | | | the search of the TOC (from the | | |
| 22 | config | ltTocLastPage | | beginning) | bool | setup |
| | | | | | | |
| 22 | config | ltTocMinEntries | | minimum number of TOC entries | bool | setup |
| | | | | ignore the tokens which are | | |
| 22 | config | spacyIgnoreBracket | | brackets ? | bool | setup |
| | | | | | | |
| | | | | ignore the tokens which are left | | |
| 22 | config | spacyIgnoreLeftPunct | | punctuation marks, e.g. "("? | bool | setup |
| | | spacyIgnoreLineTypeFoo | | ignore the tokens from line type | | |
| 22 | config | ter | | footer ? | bool | setup |
| | | spacyIgnoreLineTypeHea | | ignore the tokens from line type | | |
| 22 | config | der | | header ? | bool | setup |
| | _ | spacyIgnoreLineTypeHea | | ignore the tokens from line type | | |
| 22 | config | ding | | heading ? | bool | setup |
| | _ | spacyIgnoreLineTypeLis | | ignore the tokens from line type | | |
| 22 | config | tBullet | | bulleted list ? | bool | setup |
| | _ | spacyIgnoreLineTypeLis | | ignore the tokens from line type | | |
| 22 | config | tNumber | | numbered list ? | bool | setup |
| | 6. | spacyIgnoreLineTypeTab | | ignore the tokens from line type | | |
| 22 | config | le | | table ? | bool | setup |
| 22 | c . | | | ignore the tokens from line type | | ١. |
| 22 | config | spacyIgnoreLineTypeToc | | TOC? | bool | setup |
| 22 | c :- | T | | ignore the tokens which are | la a a l | |
| 22 | config | spacyIgnorePunct | | punctuations ? | bool | setup |
| 22 | c :- | T | | ignore the tokens which are | la a a l | |
| 22 | config | spacyIgnoreQuote | | quotation marks ? | bool | setup |
| | | | | ignore the tokens which are | | |
| 22 | config | | | right punctuation marks, e.g. ")" ? | haal | cotup |
| 22 | config | spacyighorexightrunct | | r | bool | setup |
| | | | | ignore the tokens which consist | | |
| วา | config | ana ay Tanono ano ao | | of whitespace characters ? | bool | setup |
| | COIIIIR | spacyIgnoreSpace | | ignore the tokens which are part | וטטו | setup |
| วา | config | spacyIgnoreStop | | of a "stop list"? | bool | setup |
| | config | spacyTgnorestop spacyTknAttrCluster | | brown cluster ID | bool | setup |
| | comig | Spacy inimetric ruster | | NOWII CIUSTEI ID | 5501 | Scrup |
| 22 | config | spacyTknAttrDep_ | | syntactic dependency relation | bool | setup |
| | config | spacyTknAttrDoc | | the parent document | bool | setup |
| | config | spacyTknAttrEntIob_ | | IOB code of named entity tag | bool | setup |
| | 5511116 | | | knowledge base ID that refers to | 2001 | эссар |
| | | | | the named entity this token is a | | |
| 22 | config | spacyTknAttrEntKbId_ | | part of, if any | bool | setup |
| 22 | comig | DPacy INHACCI EHCKDIA_ | I | part or, it arry | 5501 | Scrup |

Version: 0.9.8 Page: 3/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|----------|-----------|--|-----|--|------|--------|
| 22 | config | and as this nation that the first that the | | named antity type | hool | cotup |
| | config | spacyTknAttrEntType_ | | named entity type the syntactic parent, or | bool | setup |
| 22 | config | spacyTknAttrHead | | "governor", of this token | bool | setup |
| | comig | spacyimmeeriicau | | the index of the token within the | 5001 | зстар |
| 22 | config | spacyTknAttrI | | parent document | bool | setup |
| | comig | Spacy IMPREEL I | | parent document | 5001 | эстар |
| | | | | the character offset of the token | | |
| 22 | config | spacyTknAttrIdx | | within the parent document | bool | setup |
| | | | | does the token consist of | | |
| 22 | config | spacyTknAttrIsAlpha | | alphabetic characters? | bool | setup |
| | | | | does the token consist of ASCii | | |
| 22 | config | spacyTknAttrIsAscii | | characters? | bool | setup |
| | | | | | | |
| 22 | config | spacyTknAttrIsBracket | | is the token a bracket? | bool | setup |
| | | _ | | | | |
| 22 | config | spacyTknAttrIsCurrency | | is the token a currency symbol? | bool | setup |
| 22 | | | | de se the talker service of digita? | bool | |
| | config | spacyTknAttrIsDigit | | does the token consist of digits? | bool | setup |
| 22 | config | spacyTknAttrIsLeftPunc t | | is the token a left punctuation mark, e.g. "("? | bool | setup |
| | config | spacyTknAttrIsLower | | is the token in lowercase? | bool | setup |
| | comig | spacyTRIACCITSHOWCI | | is the token in lowerease: | 5001 | эстар |
| 22 | config | spacyTknAttrIsOov | | is the token out-of-vocabulary? | bool | setup |
| | config | spacyTknAttrIsPunct | | is the token punctuation? | bool | setup |
| | | | | | | |
| 22 | config | spacyTknAttrIsQuote | | is the token a quotation mark? | bool | setup |
| | | spacyTknAttrIsRightPun | | is the token a right punctuation | | |
| 22 | config | ct | | mark, e.g. ")" ? | bool | setup |
| | | | | | | |
| 22 | config | spacyTknAttrIsSentEnd | | does the token end a sentence? | bool | setup |
| 22 | config | spacyTknAttrIsSentStar | | does the token start a sentence? | hool | cotup |
| 22 | config | t | | does the token consist of | bool | setup |
| 22 | config | spacyTknAttrIsSpace | | whitespace characters? | bool | setup |
| <u> </u> | 0 | | | The state of the s | 2001 | 32.00 |
| 22 | config | spacyTknAttrIsStop | | is the token part of a "stop list"? | bool | setup |
| | config | spacyTknAttrIsTitle | | is the token in titlecase? | bool | setup |
| | config | spacyTknAttrIsUpper | | is the token in uppercase? | bool | setup |
| | | | | language of the parent | | |
| 22 | config | spacyTknAttrLang_ | | document's vocabulary | bool | setup |
| | | | | | | |
| | | | | the leftmost token of this token's | | |
| 22 | config | spacyTknAttrLeftEdge | | syntactic descendants | bool | setup |
| | c. | _, _, _ | | base form of the token, with no | | 1. |
| | config | spacyTknAttrLemma_ | | inflectional suffixes | bool | setup |
| 22 | config | spacyTknAttrLex | | the underlying lexeme | bool | setup |

Version: 0.9.8 Page: 4/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|-----------------------------|-----|---|------|--------|
| | | | | · | | |
| | | | | sequential ID of the token's | | |
| | | | | lexical type, used to index into | | |
| 22 | config | spacyTknAttrLexId | | tables, e.g. for word vectors | bool | setup |
| | | | | does the token resemble an | | |
| 22 | config | spacyTknAttrLikeEmail | | email address? | bool | setup |
| | | | | does the token represent a | | |
| 22 | config | spacyTknAttrLikeNum | | number? | bool | setup |
| | | | | | | |
| 22 | config | spacyTknAttrLikeUrl | | does the token resemble a URL? | bool | setup |
| | | | | | | |
| 22 | config | spacyTknAttrLower_ | | lowercase form of the token text | bool | setup |
| 22 | config | spacyTknAttrMorph | | morphological analysis | bool | setup |
| | | | | the token's norm, i.e. a | | |
| | | | | normalized form of the token | | |
| 22 | config | spacyTknAttrNorm_ | | text | bool | setup |
| 22 | config | spacyTknAttrOrth_ | | verbatim text content | bool | setup |
| | | | | | | |
| | | | | coarse-grained part-of-speech | | |
| 22 | config | spacyTknAttrPos_ | | from the Universal POS tag set | bool | setup |
| | | | | a length-N substring from the | | |
| 22 | config | spacyTknAttrPrefix_ | | start of the token | bool | setup |
| | | | | | | |
| | | | | smoothed log probability | | |
| 22 | config | spacyTknAttrProb | | estimate of token's word type | bool | setup |
| | | | | | | |
| | | | | sequential ID of the token's | | |
| | _ | | | lexical type, used to index into | | |
| 22 | config | spacyTknAttrRank | | tables, e.g. for word vectors | bool | setup |
| | | | | | | |
| | c. | | | the rightmost token of this | | |
| 22 | config | spacyTknAttrRightEdge | | token's syntactic descendants | bool | setup |
| 22 | t:- | | | the sentence span that this token | | |
| 22 | config | spacyTknAttrSent | | is a part of | bool | setup |
| | | | | a scalar value indicating the | | |
| 22 | config | and ar Tirn Att a Continent | | positivity or negativity of the token | haal | cotup |
| 22 | config | spacyTknAttrSentiment | | token | bool | setup |
| | | | | transform of the telepric string to | | |
| าา | config | anaavTkn\++xChana | | transform of the token's string to show orthographic features | bool | setup |
| | COIIIIR | spacyTknAttrShape_ | | length-N substring from the end | וטטו | setup |
| 22 | config | spacyTknAttrSuffix_ | | of the token | bool | setup |
| | config | spacyTknAttrTag_ | | fine-grained part-of-speech | bool | setup |
| | COTTING | pracy inimeer rag_ | +- | the token's slice of the parent | 5001 | эстар |
| 22 | config | spacyTknAttrTensor | | doc's tensor | bool | setup |
| | config | spacyTknAttrText | | verbatim text content | bool | setup |
| | 5511118 | Spacy Immeet ICAC | | text content, with trailing space | 2001 | эстар |
| | config | spacyTknAttrTextWithWs | | character if present | bool | setup |

Version: 0.9.8 Page: 5/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|------------------------|-----|-------------------------------------|-------|--------|
| | | | | the vocab object of the parent | | |
| 22 | config | spacyTknAttrVocab | | doc | bool | setup |
| | | spacyTknAttrWhitespace | | | | |
| 22 | config | _ | | trailing space character if present | bool | setup |
| 30 | font | font | х | container for font | dict | TET |
| 31 | font | embedded | | embedded | bool | TET |
| 31 | font | fontNo | | font number in document | int | TET |
| 31 | font | fullName | | font full name | str | TET |
| 31 | font | id | | font identification | str | TET |
| 31 | font | italicAngle | | font italic angle | float | TET |
| 31 | font | name | | font name | str | TET |
| 31 | font | type | | font type | str | TET |
| 31 | font | weight | | font weight | float | TET |
| 40 | heading | heading | х | container for headings | dict | LT |
| 41 | entry | ctxtLinel | | 1. context line | str | LT |
| 41 | entry | ctxtLine2 | | 2. context line | str | LT |
| 41 | entry | ctxtLine3 | | 3. context line | str | LT |
| 41 | heading | toc | Х | container for toc | dict | LT |
| 42 | heading | entries | | container for heading entries | list | LT |
| 43 | entry | entry | | container for heading entry | dict | LT |
| 43 | entry | level | | heading level | int | LT |
| 43 | entry | lineNoPage | | line number in page | int | LT |
| 43 | entry | pageNo | | page number in document | int | LT |
| 43 | entry | regexp | Х | regular expression | str | LT |
| 43 | entry | text | | heading text | str | LT |
| 50 | list | bulleted list | Х | container for bulleted list | dict | LT |
| 51 | list | format | | bullet format | str | LT |
| 51 | list | listNo | | list number in document | int | LT |
| 51 | list | noEntries | | number entries in list | int | LT |
| 51 | list | pageNoFirst | | page number first entry | int | LT |
| 51 | list | pageNoLast | | page number last entry | int | LT |
| 52 | list | entries | | container for list entries | list | LT |
| | | | | | | |
| 53 | entry | entry | | container for bulleted list entry | dict | LT |
| 53 | entry | entryNo | | entry number in list | int | LT |
| 53 | entry | lineNoPageFirst | | first line in page | int | LT |
| 53 | entry | lineNoPageLast | | last line in page | int | LT |
| 53 | entry | pageNo | | page number in document | int | LT |
| 53 | entry | paraNo | | paragraph number in page | int | LT |
| 53 | entry | text | | list entry text | str | LT |
| | list | numbered list | Х | container for numbered list | dict | LT |
| | list | format | | bullet format | str | LT |
| | list | listNo | | list number in document | int | LT |
| 61 | list | noEntries | | number entries in list | int | LT |
| | list | pageNoFirst | | page number first entry | int | LT |
| 61 | list | pageNoLast | | page number last entry | int | LT |
| 61 | list | regexp | х | regular expression | str | LT |
| 62 | list | entries | | container for list entries | list | LT |

Version: 0.9.8 Page: 6/12 Date: 28.09.2022

| lvi | container | tag | opt | description | type | source |
|----------|-----------|----------------------|-----|--|------|--------|
| 63 | entry | entry | | container for numbered list entry | dict | LT |
| _ | entry | entryNo | | entry number in list | int | LT |
| - | entry | lineNoPageFirst | | first line in page | int | LT |
| | entry | lineNoPageLast | | last line in page | int | LT |
| \vdash | entry | pageNo | | page number in document | int | LT |
| | entry | paraNo | | paragraph number in page | int | LT |
| | entry | text | | list entry text | str | LT |
| 1 | page | page | | container for page | dict | TET |
| | page | lineNoFirst | | number of first line in page | int | TET |
| | page | lineNoLast | | number of last line in page | int | TET |
| - | page | noLinesPage | | number of lines in page | int | TET |
| , _ | Page | 1102111021 430 | | manner or mice in page | | |
| 71 | page | noParasPage | | number of paragraphs in page | int | TET |
| | page | noWordsPage | | number of words in page | int | TET |
| | page | pageNo | | page number in document | int | TET |
| , _ | Page | Personal | | number of first paragraph in | | |
| 71 | page | paraNoFirst | | page | int | TET |
| , _ | Page | Paramerrise | | Page | | |
| 71 | page | paraNoLast | | number of last paragraph in page | int | TET |
| | | 1 | | | | |
| 71 | page | sentenceNoFirst | | number of first sentence in page | int | spaCy |
| 71 | page | sentenceNoLast | | number of last sentence in page | int | spaCy |
| 71 | page | wordNoFirst | | number of first word in page | int | TET |
| | page | wordNoLast | | number of last word in page | int | TET |
| 79 | page | lines | | container for lines | list | TET |
| 79 | page | paras | | container for paragraphs | list | TET |
| 80 | params | params | Х | container for parameters | dict | TET |
| 81 | params | parser | | container for module parser | dict | TT |
| 82 | params | directoryName | | file directory for intermediate files and final result files | str | params |
| 82 | params | documentId | Х | document identification | int | params |
| 82 | params | environmentVariant | | environment variant | str | params |
| 82 | params | fileNameCurr | | name of input file | str | params |
| 82 | params | fileNameNext | | name of output file | str | params |
| 82 | params | fileNameOrig | | original file name | str | params |
| 82 | params | ltHeadingRequired | | heading determination required bulleted list determination | bool | params |
| 82 | params | ltListBulletRequired | | required | bool | params |
| 82 | params | ltListNumberRequired | | numbered list determination required | bool | params |
| 82 | params | ltTableRequired | | table determination required | bool | params |
| - | params | ltTocRequired | | TOC determination required | bool | params |
| 83 | params | tokenizer | | container for module tokenizer | dict | TET |

Version: 0.9.8 Page: 7/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|------------|----------------------|-----|-----------------------------------|-------|----------------|
| 100 | line | line | | container for lines | dict | TET |
| 101 | line | lineNo | | line number in document | int | TET |
| 101 | line | lineNoPage | | line number in page | int | TET |
| 101 | line | lineNoPara | | line number in paragraph | int | TET |
| | | | | x coordinate of the lower left | | |
| 101 | line | llx | | corner | float | TET |
| 101 | line | noWordsLine | | number of words in line | int | TET |
| 101 | line | pageNo | | page number in document | int | TET |
| 101 | line | paraNo | | paragraph number in document | int | TET |
| 101 | line | paraNoPage | | paragraph number in page | int | TET |
| | | | | | | |
| 101 | line | sentenceNo | | sentence number in document | int | spaCy |
| 101 | line | tableCellNo | | cell number in row | int | TET |
| 101 | line | tableCellSpan | | cel span | int | TET |
| 101 | | tableNo | | table number in document | int | TET |
| 101 | line | tableRowNo | | row number in table | int | TET |
| 101 | line | text | | line text | str | TET |
| 101 | line | type | | line type | str | TET, TL |
| | | | | x coordinate of the upper right | | |
| 101 | line | urx | | corner | float | TET |
| 101 | line | wordNoFirst | | number word of first word in line | int | TET |
| 404 | 1. | 3 | | | | |
| 101 | line | wordNoLast | | number word of last word in line | int | TET |
| 404 | | | | word number in paragraph of | | |
| 101 | | wordNoParaFirst | | first word in this line | int | TET |
| 110 | para | para | | container for paragraph | dict | TET |
| 111 | para | lineNoFirst | | number of first line in paragraph | int | TET |
| 111 | para | IIIIeNOFIISC | | number of first line in paragraph | IIIL | IEI |
| 111 | para | lineNoLast | | number of last line in paragraph | int | TET |
| | para | noLinesPara | | number of lines in paragraph | int | TET |
| 111 | para | HOLLINESPALA | | number of files in paragraph | 1110 | 161 |
| 111 | para | noWordsPara | | number of words in paragraph | int | TET |
| | para | pageNo | | page number in document | int | TET |
| | para | pageno | | page namber in accument | | |
| 111 | para | paraNo | | paragraph number in document | int | TET |
| | para | paraNoPage | | paragraph number in page | int | TET |
| | J. 4.1. 4. | F states and a stage | | number of first sentence in | | 1 - 1 |
| 111 | para | sentenceNoFirst | | paragraph | int | spaCy |
| | | | | number of last sentence in | | , , |
| 111 | para | sentenceNoLast | | paragraph | int | spaCy |
| | para | tableCellNo | | cell number in row | int | TET |
| | para | tableCellSpan | | cel span | int | TET |
| | para | tableNo | | table number in document | int | TET |
| | para | tableRowNo | | row number in table | int | TET |
| | • | | | | | TET, |
| | | | | | | ' - ' , |

Version: 0.9.8 Page: 8/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|----------------------|-----|-----------------------------------|-------|----------|
| | | | | number of first word in | | |
| 111 | para | wordNoFirst | | paragraph | int | TET |
| | | | | number of last word in | | |
| 111 | para | wordNoLast | | paragraph | int | TET |
| 119 | para | sentences | х | container for sentences | list | spaCy |
| 119 | para | words | | container for words | list | TET |
| 120 | sentence | sentence | | container for sentences | dict | TET |
| | | | | | | |
| 121 | sentence | sentenceNo | | sentence number in document | int | TET |
| 121 | sentence | sentenceNoPage | | sentence number in page | int | TET |
| | | | | | | |
| 121 | sentence | sentenceNoPara | | sentence number in paragraph | int | TET |
| 121 | sentence | text | | sentence text | str | TET |
| | | | | number word of first word in | | |
| 121 | sentence | wordNoFirst | | sentence | int | TET |
| | | | | number word of last word in | | |
| 121 | sentence | wordNoLast | | sentence | int | TET |
| 200 | word | word | , | container for word | dict | TET |
| 201 | word | font | х | font identification | str | TET |
| 201 | word | lineNo | | line number in document | int | TET |
| 201 | word | lineNoPage | | line number in page | int | TET |
| | | | | x coordinate of the lower left | | |
| 201 | word | llx | | corner | float | TET |
| 201 | word | pageNo | | page number in document | int | TET |
| | | | | | | |
| 201 | word | paraNo | | paragraph number in document | int | TET |
| | | | | | | |
| 201 | word | sentenceNo | | sentence number in document | int | spaCy |
| 201 | word | size | х | font size | float | TET |
| 201 | word | spacyTknAttrCluster | | brown cluster ID | str | spaCy |
| | | | | | | |
| 201 | word | spacyTknAttrDep_ | | syntactic dependency relation | str | spaCy |
| 201 | word | spacyTknAttrDoc | | the parent document | str | spaCy |
| 201 | word | spacyTknAttrEntIob_ | | IOB code of named entity tag | str | spaCy |
| | | | | knowledge base ID that refers to | | |
| | | | | the named entity this token is a | | |
| 201 | word | spacyTknAttrEntKbId_ | | part of, if any | str | spaCy |
| | | | | | | |
| 201 | word | spacyTknAttrEntType_ | | named entity type | str | spaCy |
| | | | | the syntactic parent, or | | |
| 201 | word | spacyTknAttrHead | | "governor", of this token | str | spaCy |
| | | | | the index of the token within the | | |
| 201 | word | spacyTknAttrI | | parent document | str | spaCy |
| | | | | | | |
| | | | | the character offset of the token | | |
| 201 | word | spacyTknAttrIdx | | within the parent document | str | spaCy |
| | | 1 | | does the token consist of | | <u> </u> |
| 201 | word | spacyTknAttrIsAlpha | | alphabetic characters? | bool | spaCy |

Version: 0.9.8 Page: 9/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|------------------------|-----|--|----------|--------|
| | | | | does the token consist of ASCii | | |
| 201 | word | spacyTknAttrIsAscii | | characters? | bool | spaCy |
| 201 | word | spacyTknAttrIsBracket | | is the token a bracket? | bool | spaCy |
| 201 | word | spacyTknAttrIsCurrency | | is the token a currency symbol? | bool | spaCy |
| 201 | word | spacyTknAttrIsDigit | | does the token consist of digits? | bool | spaCy |
| 201 | | spacyTknAttrIsLeftPunc | | is the token a left punctuation | la a a l | |
| | word | t | | mark, e.g. "("? is the token in lowercase? | | spaCy |
| 201 | word | spacyTknAttrIsLower | | is the token in lowercase? | bool | spaCy |
| 201 | word | spacyTknAttrIsOov | | is the token out-of-vocabulary? | bool | spaCy |
| | word | spacyTknAttrIsPunct | | is the token punctuation? | bool | spaCy |
| | | 250071111100212101100 | | la circ concin parrocca are in | | opa.c, |
| 201 | word | spacyTknAttrIsQuote | | is the token a quotation mark? | bool | spaCy |
| | | spacyTknAttrIsRightPun | | is the token a right punctuation | | |
| 201 | word | ct | | mark, e.g. ")" ? | bool | spaCy |
| | | | | | | |
| 201 | word | spacyTknAttrIsSentEnd | | does the token end a sentence? | bool | spaCy |
| | | spacyTknAttrIsSentStar | | | | |
| 201 | word | t | | does the token start a sentence? | bool | spaCy |
| | | | | does the token consist of | | |
| 201 | word | spacyTknAttrIsSpace | | whitespace characters? | bool | spaCy |
| 201 | word | spacyTknAttrIsStop | | is the token part of a "stop list"? | bool | spaCy |
| | word | spacyTknAttrIsTitle | | is the token in titlecase? | bool | spaCy |
| | word | spacyTknAttrIsUpper | | is the token in uppercase? | bool | spaCy |
| 201 | word | spacytriactisopper | | language of the parent | 0001 | зрасу |
| 201 | word | spacyTknAttrLang_ | | document's vocabulary | str | spaCy |
| 201 | word | spacy in internaling_ | | document's vocabulary | 311 | зрасу |
| | | | | the leftmost token of this token's | | |
| 201 | word | spacyTknAttrLeftEdge | | syntactic descendants | str | spaCy |
| | | | | base form of the token, with no | | |
| 201 | word | spacyTknAttrLemma_ | | inflectional suffixes | str | spaCy |
| 201 | word | spacyTknAttrLex | | the underlying lexeme | str | spaCy |
| | | | | | | |
| | | | | sequential ID of the token's | | |
| | | | | lexical type, used to index into | | |
| 201 | word | spacyTknAttrLexId | | tables, e.g. for word vectors | str | spaCy |
| | | | | does the token resemble an | | |
| 201 | word | spacyTknAttrLikeEmail | | email address? | bool | spaCy |
| | | | | does the token represent a | | |
| 201 | word | spacyTknAttrLikeNum | | number? | bool | spaCy |
| | | | | | | |
| 201 | word | spacyTknAttrLikeUrl | | does the token resemble a URL? | bool | spaCy |
| 201 | | ml Title | | January State of the state of t | | |
| | word | spacyTknAttrLower_ | | lowercase form of the token text | | spaCy |
| 201 | word | spacyTknAttrMorph | | morphological analysis | str | spaCy |

Version: 0.9.8 Page: 10/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|------------------------|-----|-------------------------------------|------------|---------|
| | | | | the token's norm, i.e. a | | |
| | | | | normalized form of the token | | |
| 201 | word | spacyTknAttrNorm_ | | text | str | spaCy |
| 201 | word | spacyTknAttrOrth_ | | verbatim text content | str | spaCy |
| | | | | | | |
| | | | | coarse-grained part-of-speech | | |
| 201 | word | spacyTknAttrPos_ | | from the Universal POS tag set | str | spaCy |
| | | | | a length-N substring from the | | |
| 201 | word | spacyTknAttrPrefix_ | | start of the token | str | spaCy |
| | | | | | | |
| | | | | smoothed log probability | | |
| 201 | word | spacyTknAttrProb | | estimate of token's word type | str | spaCy |
| | | | | | | |
| | | | | sequential ID of the token's | | |
| | | | | lexical type, used to index into | | |
| 201 | word | spacyTknAttrRank | | tables, e.g. for word vectors | str | spaCy |
| | | | | | | |
| | | | | the rightmost token of this | | |
| 201 | word | spacyTknAttrRightEdge | | token's syntactic descendants | str | spaCy |
| | | | | the sentence span that this token | | |
| 201 | word | spacyTknAttrSent | | is a part of | str | spaCy |
| | | | | a scalar value indicating the | | |
| | | | | positivity or negativity of the | | |
| 201 | word | spacyTknAttrSentiment | | token | str | spaCy |
| | | | | | | |
| | | | | transform of the token's string to | | |
| 201 | word | spacyTknAttrShape_ | | show orthographic features | str | spaCy |
| | | | | length-N substring from the end | | |
| | word | spacyTknAttrSuffix_ | | of the token | str | spaCy |
| 201 | word | spacyTknAttrTag_ | | fine-grained part-of-speech | str | spaCy |
| | | | | the token's slice of the parent | | |
| | word | spacyTknAttrTensor | | doc's tensor | str | spaCy |
| 201 | word | spacyTknAttrText | | verbatim text content | str | spaCy |
| | | | | text content, with trailing space | | |
| 201 | word | spacyTknAttrTextWithWs | | character if present | str | spaCy |
| | | | | the vocab object of the parent | | |
| 201 | word | spacyTknAttrVocab | | doc | str | spaCy |
| | | spacyTknAttrWhitespace | | | | |
| - | word | _ | | trailing space character if present | | spaCy |
| | word | tableCellNo | | cell number in row | int | TET |
| | word | tableCellSpan | | cel span | int | TET |
| | word | tableNo | | table number in document | int · · | TET |
| + | word | tableRowNo | | row number in table | int | TET |
| | word | text | | word text | str | TET |
| 201 | word | type | | line type | str | TET, TL |
| | | | | x coordinate of the upper right | cı. | |
| | word | urx | | corner | float | TET |
| | word | wordNo | | word number in document | int | TET |
| 201 | word | wordNoLine | | word number in line | int | TET |

Version: 0.9.8 Page: 11/12 Date: 28.09.2022

| lvl | container | tag | opt | description | type | source |
|-----|-----------|------------|-----|--------------------------|------|--------|
| 201 | word | wordNoPage | | word number in page | int | TET |
| 201 | word | wordNoPara | | word number in paragraph | int | TET |

Version: 0.9.8 Page: 12/12 Date: 28.09.2022