Codebook Relativizers

item

Description: Running numbers listing the individual cases retrieved

file

Description: File ID giving information about the corpus (Brown/LOB/Frown/FLOB), the text category (A through R) and the individual 2,000 word chunk (two-digit number).

Retrieval: records ID (<NULL><file id=....> tag) preceding the part of the corpus the current case is retrieved from

category

Description: Text category (A through R)

Retrieval: retrieve the first character of the ID as coded in the corpus

genre

Description: Text genre of the text the case is taken from

Retrieval: Transform the text category (more fine-grained) into one of the four broad genre categories

Levels: press, generalprose, learned, fiction

rel

Description: Relativizer used in the present case

Retrieval: Looks for <DDL[SO]>, <WPR[SO]>, <W ZR> and records relativizer accordingly (for nested cases the procedure is more complicated, but the script matches each relativizer with the correct antecedent)

Levels: WHICH, THAT, ZERO

precRel

Description: Relativizer used in the construction preceding the present one

Retrieval: When new relativizer is encountered, automatically record the previous one as precrel

Levels: WHICH, THAT, ZERO

distPrecRel

Description: The distance in words between the preceding relativizer and the present one

Retrieval: When one relativizer is encountered, distPrecRel is set to 0; distPrecRel += 1 for every lexical word until the next relativizer

Calculation: lexical words between two relativizers

nested

Description: Information on whether the relativizer is part of a nested construction

Retrieval: When one (or more) relative clause construction is presently open and the script encounters a new <$.>-tag, the script registers a nested construction

Levels: yes, no

relFct

Description: Information on which syntactic function (or gap position) the relativizer assumes

Retrieval: Information extracted from the [SO] in the relativizer tags for WHICH and ZERO; THAT has only Obj.

Levels: Subj., Obj.

antPOS

Description: POS of the antecedent head

Retrieval: when antecedent head is identified (see antHead below) the word's POS information is automatically extracted

Levels: N = noun; O = other; x = unclear cases, which either need manual coding or have to be excluded

antNum:

Description: Number of the antecedent head

Retrieval: Also automatically retrieved from the antecedent head's POS-tag

Levels: 1 (for singular), 2 (plural), 0 (for cases where there is no explicit information in the POS-tag

antLN

Description: length in words between <$.>-tag and relativizer

Retrieval: When <$.>-tag is encountered, antLN is set to 0, then +=1 for every word until the relativizer

Issue: Assumes no other syntactic material between antecedent and relativizer

antGiven

Description: Givenness of antecedent head

Retrieval: When antecedent head is determined (see below, antHead), check whether the part of the text from the beginning of the 2,000 word chunk (<NULL><file id=...>tag) to the antecedent head contains the antecedent head

Levels: yes, no

antDefinite

Description: Definiteness of antecedent head

Retrieval: If there is a <APPGE>, <AT>, <DD[12]>, or <GE> tag before the antecedent head, or if the antecedent head is a <NP>, record definiteness. Else: indefinite

levels: def, indef

antFreq

Description: Text frequency of the antecedent head

Retrieval: When antecedent head is determined (see below, antHead), check how many instances of this lexical item are found in the 2,000 word chunk

Calculation: normalized for value per 1000 words

antHeadToRC

Description: Distance between antecedent head and relativizer

Retrieval: When antecedent head is determined (see below, antHead), start counting words until relativizer is encountered

RCLn

Description: Length in words of the relative clause

Retrieval: When relativizer is encountered, start counting words until <%.>-tag is encountered

LevyJaeger

Description: two words preceding and one following the relativizer (c. Levy/Jaeger 2007)

Comment: works fine for the most part, but in some cases includes tags like "<%W>" as words

TTR

Description: Type-Token ration in the 2,000 word chunk

Retrieval: for all words in the text, += 1 to tokens; add word to a list of types if it is not in that list yet; divide len(types)/tokens

meanWordLen

Description: mean word length of a given 2,000 word chunk

meanSentLen

Description: mean sentence length of a given 2,000 word chunk

stranding

Description: Proportion of stranded prepositions out of all prepositions

Retrieval: count all prepositions (<I.> tags) in 2,000 word chunk; then count the number of prepositions followed by a punctuation tag, divide the latter by the former

splitInf

Description: frequency of split infinitives
Retrieval: if elements[y] == "<TO>to" and not elements[y+1].startswith("<V")

asIfThough

Description: frequency of “as if” and “as though” in the text
Normalized per 10,000 words
Comment: intended to be a measure of adherence to a prescriptive rule, but very crude at present (ration between “as if/as though” and “like” in appropriate contexts would be better, but is difficult to retrieve automatically)

passives

Frequency of passive constructions (<VAB*> + <VVN> or <VAB*> + <RR> + <VVN>)
Normalized per 10,000 words

passiveActiveRatio

Description: proportion of passive constructions versus active lexical verbs

Retrieval:
passives: <VAB*> + <VVN> or <VAB*> + <RR> + <VVN>
actives: <VV[D0]*> or <VAH*> + <VVN>
value: passives/actives

shallWill

Description: Ratio between will and shall

Retrieval: count <VM>[Ww]ill and <VM>[Ss]hall, divide the latter by the former

nouniness

Description: amount of nouns in the text sample (as a measure of information density)

Retrieval: count all <N*>-tags

Calculation: normalized for per 10,000 words

nounVerbRatio

Description: Ratio between nouns and lexical verbs (another measure of density)

Retrieval: Count <N*>, count <VV*>, divide the former by the latter

persPronouns

Description: Amount of personal pronouns in the text sample

Retrieval: Count <PP*>-tags in the 2,000 word sample

Calculation: normalized for per 10,000 words

subordConjs

Description: Amount of subordinating conjunctions in the text sample

Retrieval: Count <CS*>-tags in the 2,000 word sample

Calculation: normalized for per 10,000 words

antHead

Description: Head of the antecedent of the relative clause in question

Retrieval: When a <$.>-tag is encountered: if <N*>, <P*>, <D*>, <M*>, or <Z*> is encountered that is not followed by <N*>, <CC>, <VVG>, or <GE>-tag: record this item as antecedent head

animacy

Description: animacy of the antecedent head
Coding: automated by running antHead across a list of the ca. 200 most frequent animate items in the corpora → hence, not a very precise measure; to be taken with a big grain of salt

adjacency

Description: Information whether the relativizer is adjacent to the antecedent head

Retrieval: not automated; manual post-editing

construction

Description: Entire construction from beginning of antecedent (<$.>-tag) to end of relative clause (<%.>-tag)

context

Description: Text context of the constructions