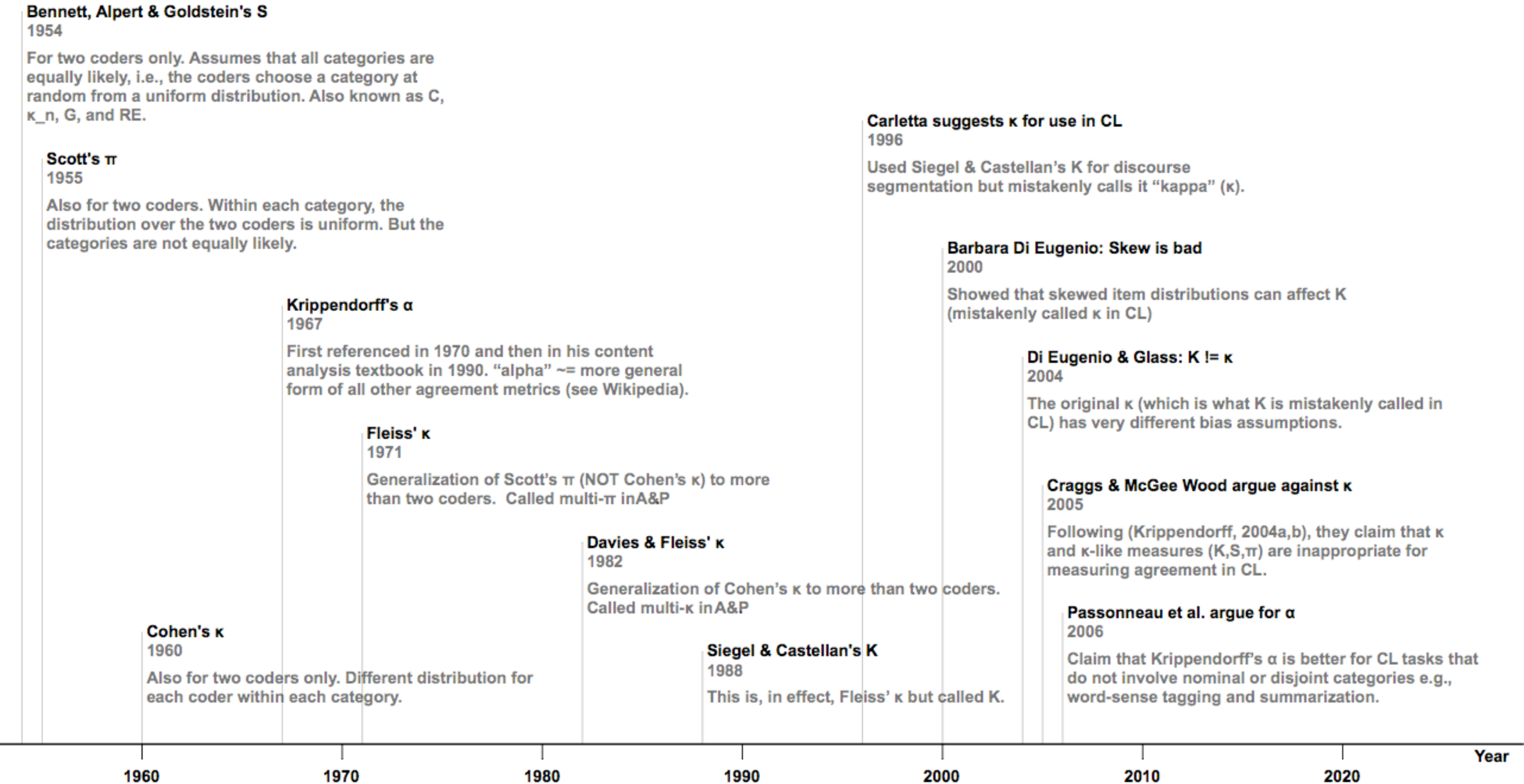
Bennett, Alpert & Goldstein's S
1954
For two coders only. Assumes that all categories are equally likely, i.e., the coders choose a category at random from a uniform distribution. Also known as C, κ_n, G, and RE.
Scott's π
1955
Also for two coders. Within each category, the distribution over the two coders is uniform. But the categories are not equally likely.
Krippendorff's α
1967
First referenced in 1970 and then in his content analysis textbook in 1990. "alpha" ~= more general form of all other agreement metrics (see Wikipedia).
Fleiss' κ
1971
Generalization of Scott's π (NOT Cohen's κ) to more than two coders. Called multi-π inA&P
Davies & Fleiss' κ
1982
Generalization of Cohen's κ to more than two coders. Called multi-κ inA&P
Cohen's κ
1960
Also for two coders only. Different distribution for each coder within each category.
Siegel & Castellan's K
1988
This is, in effect, Fleiss' κ but called K.
Carletta suggests κ for use in CL
1996
Used Siegel & Castellan's K for discourse segmentation but mistakenly calls it "kappa" (κ).
Barbara Di Eugenio: Skew is bad
2000
Showed that skewed item distributions can affect K (mistakenly called κ in CL)
Di Eugenio & Glass: K != κ
2004
The original κ (which is what K is mistakenly called in CL) has very different bias assumptions.
Craggs & McGee Wood argue against κ
2005
Following (Krippendorff, 2004a,b), they claim that κ and κ-like measures (K,S,π) are inappropriate for measuring agreement in CL.
Passonneau et al. argue for α
2006
Claim that Krippendorff's α is better for CL tasks that do not involve nominal or disjoint categories e.g., word-sense tagging and summarization.
1960
1970
1980
1990
2000
2010
2020
Year