Supplementary Material

When Do Politicians Use Populist Rhetoric? Populism as a Campaign Gamble

Yaoyao Dai & Alexander Kustov

Online Appendix A:

Model of Populism as a Campaign Gamble

Building on Skaperdas and Grofman (1995), our model assumes a standard office-seeking two-candidate race where each candidate has a certain level of pre-existing electorate support among the decided (or henceforth "mobilized") voters, while the rest of the (potential) electorate is undecided (or henceforth "unmobilized"). We further stipulate that candidates can use a combination of "conventional" and "populist" campaign rhetoric to improve their electoral chances. While conventional rhetoric is assumed to help mobilize additional support among the unmobilized, populist rhetoric is assumed to demobilize the opponent's pre-existing support. The use of populism, however, can backfire such that there is some chance that it can demobilize the candidate's own pre-existing support (or, equivalently, mobilize more votes for the other candidate). Overall, we show that, despite the potential risk, the *ex-ante* losing candidate is more likely to use populism to have at least some chance of winning.

Our basic model of populist rhetoric as a campaign gamble is of imperfect information. There are two candidates (or parties), A and B. Both candidates observe each others' level of pre-existing support α_i and the share of unmobilized electorate $\overline{\alpha} = 1 - \alpha_A + \alpha_B$. Then, each candidate simultaneously decides to allocate its effort to populist (p_i) or non-populist, conventional (c_i) rhetoric so that $p_i + c_i = 1$.

While political campaigning can have a number of aims including changing voter's preferences over candidates, we assume that the primary function of conventional campaign rhetoric is mobilizing electoral support among the (currently) unmobilized. Put formally, let $m^A(c_A, c_B)$ and $m^B(c_A, c_B)$ indicate the share of unmobilized electorate ultimately attracted by candidates A and B such that, for any given combination pair of conventional campaign strategies pursued by both candidates, $m^i(c_i, c_j)$ is increasing in c_i and decreasing in c_j. To that end, we also assume that all of the unmobilized

are equally and ultimately susceptible to mobilization by either candidate such that $m^A(c_A, c_B) = m^B(c_A, c_B)$ and $m^A(c_A, c_B) + m^B(c_A, c_B) = 1$. As a result, the candidates attract the same share of the unmobilized electorate when they decide to allocate the same amount of effort to conventional campaigning. Since the function m is symmetric ($m^i(c_i, c_j) = 1 - m^j(c_j, c_i)$), we can simply denote $m^A(c_A, c_B)$ by m and $m^B(c_A, c_B)$ by $1 - m$. Finally, we assume that conventional campaigning has diminishing returns (so that m is a concave function: $d^2m/dc_i^2 \leq 0$).

Unlike conventional campaigning to attract the unmobilized, we assume that the primary function of populist rhetoric is demobilizing the opponent's pre-existing support. In line with some of the empirical literature described above, however, we also assume that populist campaigning can backfire by demobilizing the candidate's own current supporters. Put formally, for any given combination pair of populist campaign strategies p_i pursued by both candidates, let $\alpha_i(p_i + Ep_j)$ be the resulting decrease in pre-existing support share for candidate i, where E indicates the relative effectiveness of populist campaigning (or to what extent the opponent is hurt more than the candidate). In our base model, we assume that $E > 1$. Importantly, at least in terms of the relative electoral advantage, the backfire effect of populist campaigning that demobilizes one's own support is equivalent to the one that mobilizes the electorate to vote for the opponent. In other words, although the model focuses on populism as primarily a tool for demobilization, the main distinctive feature of populist rhetoric is ultimately assumed to be its greater riskiness (relative to non-populist rhetoric).

We can now summarize the final overall support that each candidate gets after deciding on their use of conventional and populist rhetoric. To simplify, given that $p_i + c_i = 1$, we can represent the resulting support (α_i') as just a function of c_i:

$$\alpha_i' = \alpha_i + \overline{\alpha}m - \alpha_i(1 - c_i + E(1 - c_j)) \tag{1}$$

Similar to other campaign strategy literature, we assume that candidates ultimately care about maximizing their winning margin or electoral advantage. Consequently, given equation 1, we can define the utility function for each candidate as follows:

$$u_i(c_i, c_j) = \alpha_i' - \alpha_j' = \overline{\alpha}(2m(c_i, c_j) - 1) - \alpha_i(1 - c_i + E(1 - c_j)) + \alpha_j(1 - c_j + E(1 - c_i)) + \alpha_i - \alpha_j \quad (2)$$

After formulating the strategic form of our game, we can now proceed with determining the possible Nash equilibria. We can say that a campaign strategy pair (c_A^*, c_B^*) is an equilibrium if $u_i(c_i^*, c_j^*) \geq u_i(c_i, c_j^*)$ for all $c_i, i \neq j$. Let $u_i' = du_i(c_i, c_j)/dc_i$, $m_i' = dm/dc_i$, and assume that $m_i'' = d^2m/dc_i^2 \leq 0$. Then, we can find the first derivative and characterize the marginal benefits of putting extra effort into conventional and populist campaigning as follows:

$$u_i'(c_i, c_j) = \overline{\alpha}2m_i' - (E\alpha_j - \alpha_i). \quad (3)$$

We can then similarly derive $u_i'' = \overline{\alpha}2m_i'' \leq 0$. Conditional on the assumptions above being satisfied, we can now show that candidates' utility function u_i is concave in their own strategy c_i and thus that there exists a pure-strategy Nash equilibrium.

Given equation 3, both candidates would only devote a non-zero effort to both campaign strategies ($0 < c_A^* < 1$ and $0 < c_B^* < 1$) if and only if their marginal benefits and costs are equalized. Quite naturally, this implies that $E\alpha_A > \alpha_B$ and $E\alpha_B > \alpha_A$ simultaneously, which necessarily requires that the derivatives in equation 3 are equal to zero for both candidates:

$$\begin{aligned} \overline{\alpha}2m_A'(c_A^*, c_B^*) - (E\alpha_B - \alpha_A) = 0 &\implies m_A'(c_A^*, c_B^*) = (E\alpha_B - \alpha_A)/(2\overline{\alpha}) \\ -\overline{\alpha}2m_B'(c_A^*, c_B^*) - (E\alpha_A - \alpha_B) = 0 &\implies -m_B'(c_A^*, c_B^*) = (E\alpha_A - \alpha_B)/(2\overline{\alpha}) \end{aligned} \quad (4)$$

Now suppose that one of the candidates has more pre-existing support $\alpha_A > \alpha_B$.

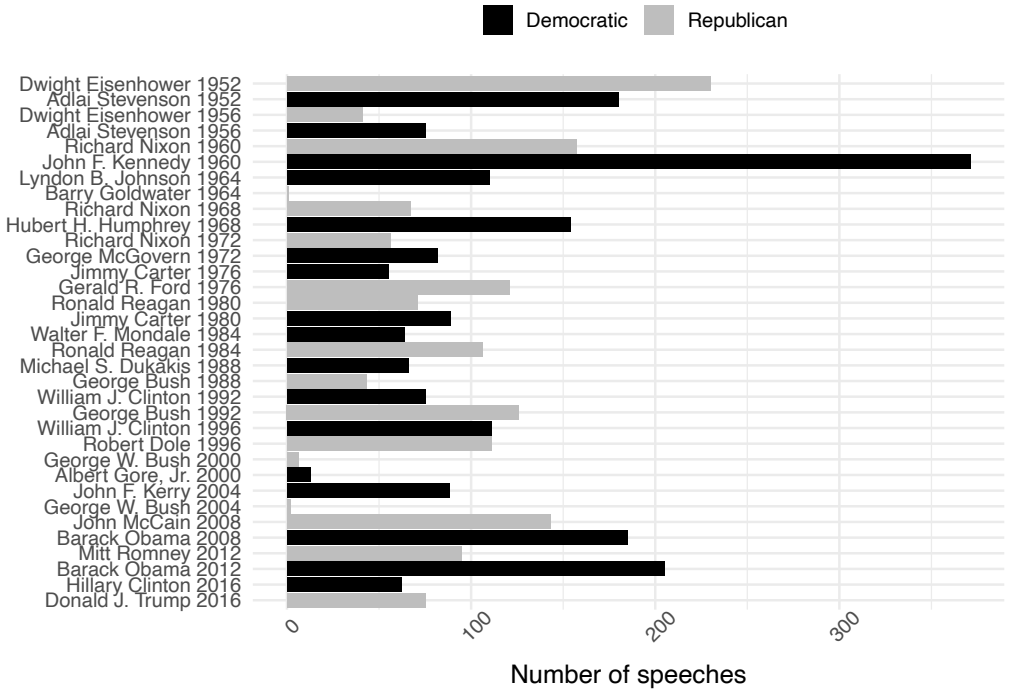
We can then show that $E\alpha_A - \alpha_B > E\alpha_B - \alpha_A$ and, by equation above, $m'_A(c^*_A, c^*_B) < -m'_B(c^*_A, c^*_B)$ so that $m(c^*_A, c^*_B) > 1/2$ under our assumptions. In turn, this is equivalent to $c^*_A > c^*_B$ or $p^*_A < p^*_B$, which gives us our main result: *"the candidate with a lower pre-existing support is expected to use more populist campaign rhetoric relative to his opponent"* (**Proposition 1**).

In addition to this general result it may also be instructive to examine two special cases where one of the candidate allocates all effort to either conventional or populist campaigning (c^*_i is equal to 0 or 1). First, suppose that the pre-existing support is lower for one of the candidates ($\alpha_i > \alpha_j$) and that the effectiveness of populist rhetoric is relatively low ($E\alpha_j \leq \alpha_i$). Then, in line with equation 3, $u'_i(c_i, c_j) > 0$ given that $\alpha_i \geq E\alpha_j$. Consequently, candidate i would only do conventional campaigning in equilibrium ($c^*_i = 1$ is the optimal choice regardless of c_j). Second, consider a function m with a finite derivative $m'_i(0, c_j)$ and sufficiently low α_i (or sufficiently high E). Then, regardless of c_j, it must be true that $u_i(0, c_j) = \overline{\alpha} 2m'_i(0, c_j) - (E\alpha_j - \alpha_i) \leq 0$. In other words, candidate i would only do populist campaigning in equilibrium ($c^*_i = 0$ is the optimal choice regardless of c_j). In sum, although this is less realistic than the general proposition 1, *if the candidate's pre-existing support is sufficiently low (high) or populist rhetoric is sufficiently (in)effective, then the candidate is expected to fully engage in populist (conventional) campaigning.* Importantly, the results hold even if we introduce some uncertainty about the (in)effectiveness of populism and relax the assumption that $E > 1$, i.e., that populist rhetoric is necessarily hurting the opponent more than the candidate instigator of such rhetoric (not shown).

Online Appendix B:

Tables and Figures

Figure B1: Distribution of Speeches across Presidential Campaigns



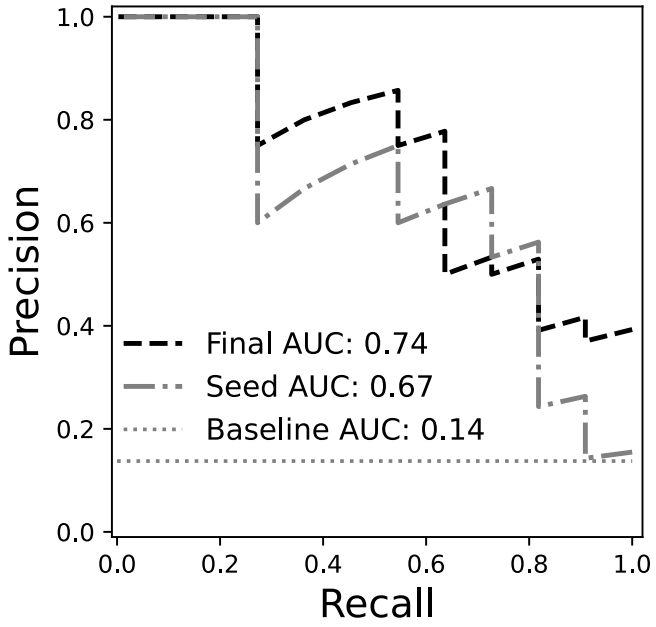
Note: The plot visualizes the distribution of speeches across campaigns in our combined dataset.

Table B1: Summary Statistics

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Populism Rhetoric (Average Share)	0.02	0.09	0	0	0	1
Electoral Advantage (Binary)	0.53	0.50	0	0	1	1
Electoral Advantage (Percent)	0.44	6.40	−29.50	−4.00	5.00	18.00
Party Incumbency	0.46	0.50	0	0	1	1
Party Membership (Republican)	0.42	0.49	0	0	1	1
Speech Length (Standardized)	0.19	0.12	0.00	0.10	0.26	1.00

Full Speech-level Data (n = 3,436). The mean speech length is 2,167 words.

Figure B2: Out-of-sample Model Performance before and after Active Learning



Note: The plot indicates the precision recall AUCs for the random forest classifier using the seed training data and the final model after populating the training data with active learning.

Table B2: Populist Rhetoric as a Function of Electoral Advantage (Robustness Checks)

	Sub-speech Level	Speech Level	
	(1)	(2)	(3)
Electoral Advantage	-0.013^{***}	-0.015^{**}	-0.015^{**}
	(0.003)	(0.005)	(0.005)
Party Incumbency	-0.039^{***}	-0.003	-0.003
	(0.003)	(0.008)	(0.007)
Partisanship (GOP)	0.005	0.015	0.014
	(0.003)	(0.015)	(0.013)
Speech Length	N/A	Yes	Yes
Month FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Candidate FE	No	Yes	Yes
Excluding D. Trump	No	No	Yes
Observations	11,839	3,436	3,361
Adjusted R^2	0.131	0.322	0.136

Note: All models are OLS regressions of populist rhetoric in U.S. presidential speeches on electoral advantage in the recent monthly polls and other candidate or speech characteristics. Standard errors are given in parentheses, $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

Table B3: Populist Rhetoric as a Function of Electoral Advantage (Sep.-Nov. Only)

	(1)	(2)	(3)	(4)	(5)
Electoral Advantage	-0.013^{***}	-0.013^{***}	-0.012^{***}	-0.021	-0.020^{*}
	(0.003)	(0.003)	(0.003)	(0.011)	(0.009)
Party Incumbency	-0.021^{***}	-0.023^{***}	-0.027^{***}	-0.004	-0.003
	(0.003)	(0.003)	(0.003)	(0.007)	(0.006)
Partisanship (GOP)	0.002	0.002	0.00004	0.023	0.021
	(0.003)	(0.003)	(0.003)	(0.017)	(0.014)
Speech Length	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Month FE	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Year FE	*No*	*No*	*Yes*	*Yes*	*Yes*
Candidate FE	*No*	*No*	*No*	*Yes*	*Yes*
Excluding D. Trump	*No*	*No*	*No*	*No*	*Yes*
Observations	2,778	2,778	2,778	2,778	2,725
Adjusted R^2	0.021	0.024	0.274	0.433	0.246

All models are OLS regressions of the average share of populist rhetoric in U.S. presidential speeches on electoral advantage in the three last monthly polls and other characteristics.
The standard errors are given in parentheses, *p<0.05; **p<0.01; ***p<0.001.

Table B4: Populist Rhetoric as a Function of Electoral Advantage (Candidate-Year-Month Level)

	(1)	(2)
Electoral Advantage	-0.009	-0.011
	(0.009)	(0.009)
Month FE	*No*	*Yes*
Year FE	*No*	*Yes*
Observations	189	189

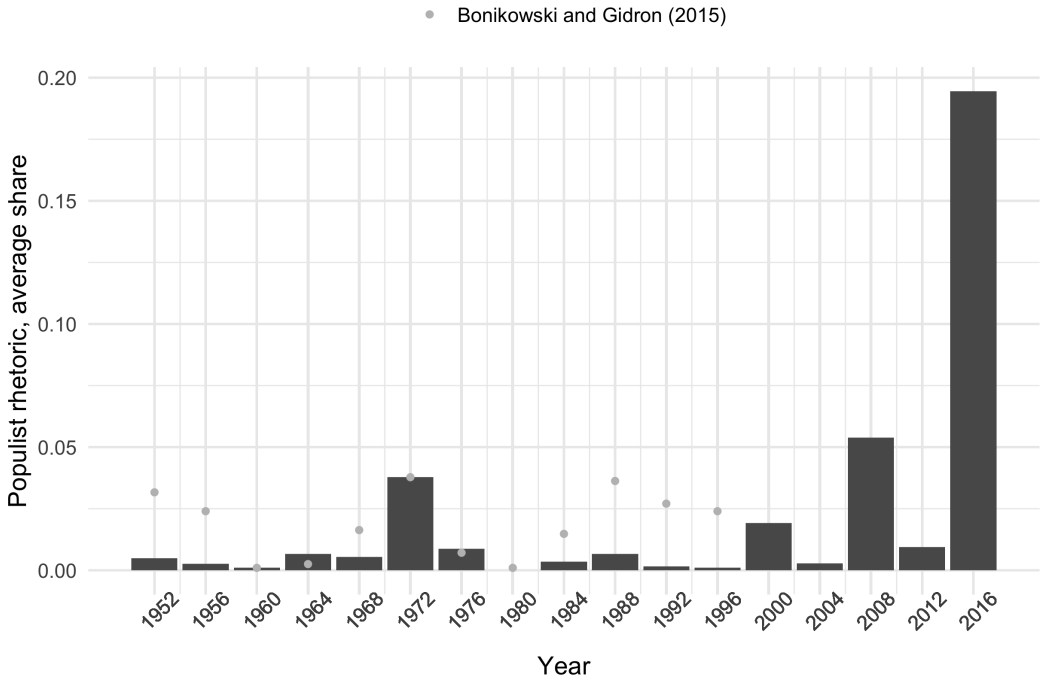
All models are OLS regressions of the average share of populist rhetoric among U.S. presidential candidates on electoral advantage in the most recent monthly polls aggregated at the candidate-year-month level. The standard errors are given in parentheses.

Table B5: Populist Rhetoric as a Function of Electoral Advantage (Rare Event Models)

	Negative Binomial				*Zero-inflated Negative Binomial*			
	(1)	(2)	(3)	(4)	(5a)	(5b)	(6a)	(6b)
Electoral Advantage	-0.860^{***}	-0.397^{*}	-0.371	-0.544	-1.865^{***}	-0.610	-1.216^{***}	-4.114^{***}
	(0.197)	(0.202)	(0.231)	(0.290)	(0.296)	(0.341)	(0.298)	(1.042)
Candidate Controls	*Yes*	*Yes*	*Yes*	*Yes*	*No*	*No*	*Yes*	*Yes*
Sub-Speech Count	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*	*Yes*	*No*
Month FE	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*	*Yes*	*Yes*
Year FE	*No*	*No*	*Yes*	*Yes*	*No*	*No*	*Yes*	*Yes*
Candidate FE	*No*	*No*	*No*	*Yes*	*No*	*No*	*No*	*No*
Observations	3,436	3,436	3,436	3,436	3,436	3,436	3,436	3,436
Log Likelihood	-824	-734	-569	-539	-850	-850	-502	-502

All models are negative binomial regressions of the populist sub-speech count in the speeches of U.S. presidential candidates. The last four columns show the two-part zero-inflated negative binomial regressions (5 and 6) with the first (a) column indicating the count model coefficients and the second (b) column indicating the zero-inflation model coefficients. The standard errors are given in parentheses. *p<0.05; **p<0.01; ***p<0.001.

Figure B3: Average Share of Populist Rhetoric across Years



Note: The bar plot indicates the average share of populist sub-speeches in U.S. presidential campaigns across years based on our measure and Bonikowski and Gidron (2016). Since we have a sub-speech level measure and Bonikowski and Gidron (2016) measure is at word level, the two measures are on different scales. To make it more comparable, their measure is re-scaled to have the same lower and upper limit as our measure between 1952 to 1996. To match and to compare with Bonikowski and Gidron (2016), we create the annual average level of populist rhetoric by calculating the proportion of populist sub-speeches in a year. As can be seen, both measures have similar trends between 1952 and 1996 with 1972 and 1988 as the most populist elections while 1960 and 1980 as the least populist elections.

Online Appendix C:

Word Embedding Models

Word embedding is a type of language model that maps words or sentences and documents into vectors of real numbers. Unlike the common 'bag-of-words' method of vectorization, in which one unique word is one dimension, word embedding represents words and documents in a dense continuous vector space with many fewer dimensions and positions semantically and syntactically similar words close to each other in this vector space. The method of word embedding is based on a distributional hypothesis in linguistics theory, which states that the meaning of a word is a function of its contexts or surrounding words. Unlike the 'bag-of-words' assumption, which treats words as independent atomic units, the distributional hypothesis aims to model the meaning of a word and assumes that the meaning of a word is given, and can be approximated, by the sets of contexts in which the word appears. In effect, the underlying idea is that words that frequently appear in the same contexts are likely to have a similar meaning.

There are several different ways to train word embedding. In this paper, we use a Doc2vec model (Le and Mikolov, 2014), which is based on the more foundational Word2vec model (Mikolov et al., 2013). We begin by describing the Word2vec model. The Word2vec model is a neural network based model that takes each unique word in the vocabulary of a corpus as an input. The input word, represented as a one-hot vector, is then multiplied by a dense, real-valued weights matrix of size $V \times d$, where V is the length of the vocabulary in the corpus and d is the chosen size of the hidden layer or 'embedding'.[1] By multiplying the $1 \times V$ input vector for a word with the $V \times d$ weights matrix, a $1 \times d$ vector is generated; this is the word's vector representation, v_{word}. The model then uses this vector representation of the input target word as the input to a softmax classifier to predict which of the V words in the vocabulary

[1] We choose $d = 150$, in keeping with standard practice.

are likely to be the context words of the input word. Context words are those that appear in a certain range of words before and after the current/target word. The model learns the embedding or the parameters in the hidden layer by finding the parameters that maximize the predicted probability of true context words. In other words, the Word2vec model seeks to set parameters θ to maximize the conditional probability of contexts C when observing the target word T: $p(C|T;\theta)$ for all words in the vocabulary (Mikolov et al., 2013; Goldberg and Levy, 2014).[2] Therefore, mathematically, the model assigns similar parameters to words that are used interchangeably in the same contexts.

Because maximizing $p(C|T;\theta)$ for all target and possible contexts is expensive to compute and there are more words that do not appear together than words that often appear together, we adopt negative sampling skip-gram in training the model. In negative sampling skip-gram, the input layer contains target-context word pairs. The target-context pairs are generated by taking the target word at index i and pairing it with all context words from $i-k$ to $i+k$ given a window size k.[3] For every true target-context word pair, we generate s negative samples; these are target-context word pairs that are not observed in the actual text corpus.[4] The output layer contains dummy values 1 and 0 indicating whether the input pair is a true target-context pair that co-locates in the texts (1) or a negative/fake pair that does not appear together in the texts (0). The predicted value given an input pair is computed by taking the dot product of the target word vector (target embedding) and the context word vector (context embedding) and then applying the logistic function, $\sigma(\cdot)$. The model uses small non-zero random values as the initial parameters in the hidden layer to produce

[2]Word2vec encompasses two different, related neural-network based models, including the continuous bag-of-words (CBOW) and skip-gram (SG) models (Mikolov et al., 2013). The SG model, which is used and explained here, inputs a target word from a text and attempts to predict the target word's likely context words. The CBOW model does the reverse. Given a set of context words, CBOW attempts to predict the context's target word.

[3]In our model we use $k = 10$.

[4]In our model, we use $s = 10$.

the embedding/word vector. Stochastic gradient descent is then used to optimize the parameters through back-propagation to minimize the logarithmic loss between $\sigma(v_{\text{target word}} \cdot v_{\text{context word}})$ and the true value $[0, 1]$.

Expanding the Word2vec model to the document level is simple; each document is labeled with an ID and treated as one unit (like a word). This document ID is positioned within the text in the document. For example, suppose we have a one-sentence document labeled as Doc1: "We are fighting for the forgotten Americans." The document ID is treated as one unit and positioned within its text: "We are fighting for *Doc1* the forgotten Americans."[5] The negative sampling algorithm can now be applied to both the target word and the document, which is treated as a target word. In this way, the documents sharing similar texts or content are positioned close to each other in the vector space (Le and Mikolov, 2014).

[5]In practice, the model is adjusted so that the *Doc1* token occurs in all of document 1's words' contexts and all of the document 1's words appear in the *Doc1* token's context.