自然言語処理 110－3
（1995．11．17）

品詞の並びに関するヒューリスティックスを用いた日本語同語反復表現の検出

滝澤修　井佐原均

郵政省通信総合研究所関西先端研究センター

自然言語における修辞的表現の一種である「同語反復表現」を計算機で検出する一手法を提案する．同語反復表現とは，「彼は彼，私は私だ」や「建物という建物が倒壊した」のように，一文中で同じ語（反復語）が繰り返され，かつ表層的な処理では意味解析できない表現のことである．本稿で提案する手法は，形態素解析によって品詞の同定を行い，反復語とその周辺の単語の品詞の並びに関するテンプレートを学習データから予め人手で作成しておき，そのテンプレートとのマッチングによって，対象とする文から同語反復表現を検出するものである．新聞記事データを用いた予備的な実験では，本手法によって，80.6%の呼出率で検出できることが示された．

A Method for Mechanically Detecting Japanese Tautological Expressions Using Template Matching of POS

Osamu Takizawa and Hitoshi Isahara

Kansai Advanced Research Center,
Communications Research Laboratory,
Ministry of Posts and Telecommunications
{taki, isahara}@crl.go.jp

This report proposes a method for mechanically detecting tautological expressions in the Japanese language. Tautological expressions are defined here as "rhetorics in which same words are appeared at two positions in one sentence", e.g., "A promise is a promise". In the proposed method, tautological expressions are detected by matching with templates; each template is a chain of POS (part of speech) of repeated words and the neighbouring words. The templates are created manually beforehand. Preliminary experiments using a newspaper corpus show the appropriateness of the proposed method.

1．はじめに

機械翻訳などの自然言語処理システムにおいて，表層的な処理では意味解析できない表現を予め検出して別処理ルーチンへ迂回させる方法を用いることは，効率的な処理のために重要である．そのような表現の一つとして，同語反復表現がある．例えば「建物という建物が倒壊した」のような定型的表現は，表層的に構造を分析しただけでは，正しく意味解釈ができない．また「約束は約束だ」のようなトートロジー表現[1]は，反復語に関する一部の属性についての推論を起動しないと意味解釈ができない．

従来から行われている定型表現の検出の研究は，述語型定型表現（例：「手を染める」）のように，表現全体の語彙を限定して（例えばこの例の場合は「AをB」の形式でかつ

「A＝手」「B＝染める」に限定される）ひとまとまりのものとして扱える表現を対象とするものがほとんどであった（そのために，例えば「手」と「染める」の共起関係を使った検出等が試みられている[2]）．それに対して同語反復表現は，一般に反復語に自由度がある（つまり反復語を例えば同じ品詞の他の語に入れ換えても，計算機処理上は同種の表現として扱える場合が多い）ため，語まで限定されたひとまとまりの定型表現として扱うことはできない．同語反復表現のように，構成語に自由度がある定型表現を検出する研究は，「あやうく…するところだった」のような多語性慣用表現の他には，まだあまりなされていない．

本稿では，同語反復表現を計算機で検出する一手法を提案する．本稿で提案する手法は，形態素解析によって品詞の同定を行い，反復語とその周辺の単語の品詞の並びに関するテンプレートを学習データから予め人手で作成しておき，そのテンプレートとのマッチングによって，対象とする文から同語反復表現を検出するものである．

小規模コーパス（新聞記事データ）を用いた予備的な実験では，本手法によって，80.6％の呼出率で検出できることが示された．

2．同語反復表現についての検討

2.1 用語の定義

同語反復表現における反復語のうち，文中で先に現れるものを「前反復語」，後のものを「後反復語」と呼ぶことにする．また，両反復語の間の語列のことを「反復語間語列」と呼ぶことにする．さらに，後反復語の後に続く語列のことを「後続語列」と呼ぶことにする．例えば「建物という建物が倒壊した．」の場合，前反復語と後反復語は共に「建物」，反復語間語列は「という」，後続語列は「が倒壊した．」となる．

本稿において扱う同語反復表現は，以下の3つの条件をすべて満たしているものとする

（1）反復語は，基本形（原形）が同じ自立語であること．

（2）反復語間語列の語数は1以上であること．

従って「みるみる」「たびたび」「ますます」「よくよく（考える）」のように，反復語間語列がない表現は，本稿では同語反復表現とはみなさない．このような表現は一つの語彙として辞書登録しておくべき表現である場合が多く，自然言語処理システムにおいて，予め検出して別処理ルーチンへ迂回させる必要性が少ないと思われる．

（3）表層的な処理だけでは，意味解釈ができない表現であること．

例えば「建物という建物が倒壊した」のような定型的表現は，表層的な構造分析だけを行うと，例えば「『建物』と命名されている建物が倒壊した」のような誤った意味解釈に陥る可能性がある．また「約束は約束だ」のようなトートロジー表現は，例えば「約束は必ず履行しなければならないという属性を持っている」のように意味解釈する必要があり，そのためには反復語「約束」に関する一部の属性についての推論を起動しなければならない．従ってこれらは表層的な処理だけでは意味解釈ができない表現といえるので，本研究における同語反復表現とみなす．

それに対して，例えば「憎悪が憎悪を呼ぶ」「言うべきことを言う」「難航に難航を重ねる」などは，同じ語が反復されている定型的な表現ではあるが，表層的な構造の分析だけで意味解釈ができ，意味はその分析によって得られる以上の意味を持たないと考えられる．従って本研究における同語反復表現とはみなさない．

2.2 検出機構の方針

定型表現の検出は，字面処理のみで統計的手法を用いた浅い処理によるのが最近の傾向である[3]．しかし同語反復表現の場合は，字面処理のみかつ統計的手法を使うのは困難である．その理由として，大きく以下の3つが挙げられる．

（1）同語反復表現はコーパス中での出現頻度があまり高くない．そのため統計的手法を用いても有意な結果が得られることは期待できない．

（2）「送るだけ送って下さい」のように，反復語の出現形は付属部が屈折している場合

があるため，反復語の検出のためには基本形に戻して照合する必要がある．そのため字面処理のみでは対応困難である．

（３）反復語になり得るのは，その語だけで意味をもつ語，すなわち自立語という制約があるため，この制約を検出に活用するためには，語の品詞の同定が必要である．

しかし同語反復表現の検出は，あくまでも自然言語処理における前処理的な位置づけであるので，構文解析や意味解析のような重い処理を用いることは非効率的な上，誤った処理結果によって後処理にかえって悪影響を及ぼす恐れがある．そのため，できるだけ簡易で誤りの少ない処理で済ませることが望ましい．

以上より，本稿で提案する手法では，自然言語処理としては形態素解析だけを行い，単語の基本形と品詞を同定して，品詞の並びに関するテンプレートマッチングによって検出する方法を採用する．具体的には，実際のコーパスを学習データとし，同語反復表現であると検査者が判断した表現を抽出して，その反復語，反復語間語列および後続語列の品詞の並びに関するテンプレートを人手で作成しておく．そして，そのテンプレートを処理対象のデータと照合し，マッチしたものを検出結果として出力する．この手法を用いると，品詞並びパターンの分類も同時にできるため，パターン毎に後処理が異なる場合にも対応が容易なので，都合がよい．

３．検出処理の概要

まず，学習データを用いて人手によってテンプレートを作成しておく．そしてそのテンプレートを用いて，対象とする文書を処理する．

３．１　人手によるテンプレートの作成

人手によるテンプレート作成において，大量の学習データをすべて人手で検査して同語反復表現を抽出することは，労力や信頼性等の観点から現実的でない．そこで，同語反復表現が最低限満たすべき制約を用いて絞り込みを行い，絞り込んだ結果だけを人手検査の対象とする．図１にテンプレート作成のための処理の流れを示す．

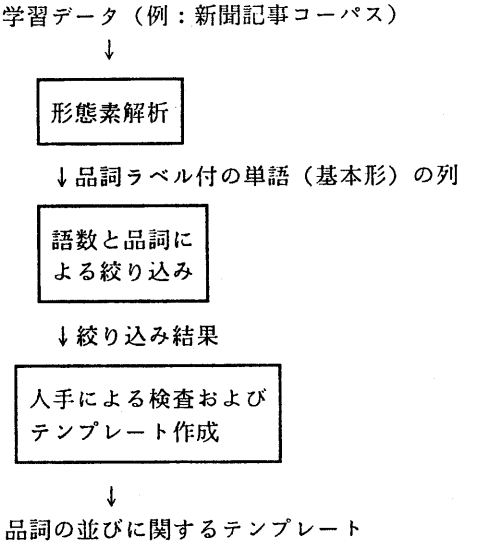

図１　テンプレート作成のための処理

この処理ではまず最初に形態素解析を行い，単語に分割して品詞のラベルを付与する．形態素解析には(株)リコーの簡易日本語解析系Q_JPを用いた[4]．そのため，本研究における品詞分類も同解析系に依存している．同解析系における品詞分類は，大品詞，中品詞，小品詞の３レベルがある．例えば大品詞は「助詞」，中品詞は「格助詞」，小品詞は「格助詞ノ」のようになる．小品詞は，助詞や助動詞等についてのみ定義されている分類であり，「中品詞＋基本形」の形式で定義されている．形態素解析によって，各単語の基本形とその中品詞または小品詞が得られる．

次に，語数と品詞による絞り込みを行う．すなわち，次の条件を満たす単語列を選別する．

「単語の並びが，『単語A-単語1-(単語2-(単語3-(単語4-(単語5))))-単語B』となっている（括弧内は省略可能）．但し単語Aと単語Bは，基本形と品詞が共に一致しており，かつその品詞は，動詞，形容詞，名詞の一部（普通名詞，形式名詞，指示名詞，相対名詞），サ変名詞のいずれかである．単語1～5は任意とする．」

「世界を市場とし、植民地としたことによる...」

↓

	単語A	単語1	単語2	単語3	単語B
抽出単語列	し	、	植民地	と	し
基本形	する	、	植民地	と	する
品詞	サ変動詞	読点	普通名詞	格助詞ト	サ変動詞

図2　語数と品詞による絞り込みによって抽出された単語列の例

すなわちこの絞り込みは，単語A（基本形は単語Bの基本形と一致）の品詞の制限と，単語A－B間の語数の制限（１～５個）を行うものである．この絞り込みによって抽出された単語列の例を図２に示す．単語A－B間の語数の上限を５個に設定したのは，学習データ（朝日新聞1994年社説１年分）[5]に対する人手による検査において単語A－B間の語数を色々変えて試した結果，反復語間語列の語数が６個以上の同語反復表現が全く無かったことに基づく．例えば「不正は，もちろん誰が何と言おうと間違いなく絶対に不正だ」のように，単語A－B間の語数が６個以上のものももちろん原理的にはありえるが，実際のコーパスにはまず出現しないという経験的仮定に立って制限している．

次に，絞り込み結果を人手で調べ，同語反復表現と判断されたものについて，その反復語（単語A，B），反復語間語列（単語1～5），および後続語列の品詞の並びをテンプレートとする．但し後続語列については，同語反復表現の構成にかかわると判断される範囲のみをテンプレートに入れる．

テンプレートとしては，学習データから抽出された中品詞または小品詞をそのまま用いたもの（以下「厳しい制約によるテンプレート」）と，品詞制限を緩和したもの（以下「緩い制約によるテンプレート」）との２種について試してみた．

（１）厳しい制約によるテンプレート

朝日新聞1994年社説１年分(22,385文)[5]を学習データとして作成した，厳しい制約によるテンプレートを表１に示す．表の横スレッドが一つのテンプレート（反復語，反復語間語列および後続語列の品詞の並び）になっており，複数の品詞が縦に書かれているブロックは，その品詞のうちのどれでもよいことを意味する．例えば，表１のNo.5のテンプレートを用いると，「倒産するべくして倒産する」がマッチする．

なお実際のテンプレートは，前反復語-反復語間語列-後反復語-後続語列，の順に並んだ品詞列であるが，前反復語と後反復語とは常に同じなので，表１では両者をまとめて「反復語」として掲げている．

（２）緩い制約によるテンプレート

検出機構において，検出漏れはできるだけ避けるべきエラーである．検出漏れを減らす方法の一つとして，制約を緩くしたテンプレートを用いることが考えられる．但し，制約を緩くすることによって，検出漏れを減らせる代わりに誤検出が増える恐れがある．従って制約緩和基準は慎重に決める必要がある．

本稿では，品詞に関する以下の緩和基準を用いて，緩い制約によるテンプレートを作成した．以下の緩和基準に従って表１を緩和して得られた，緩い制約によるテンプレートを表２に示す．表２のテンプレートのNo.は，表１のNo.と対応するようにしている．

【テンプレートの制約緩和基準】

制約の緩め方は，３ランクある品詞体系の直近上位の品詞（小品詞ならば中品詞，中品詞ならば大品詞）に拡張することを基本とする．

［反復語］

中品詞を大品詞に拡張する．更に動詞（形容詞）の場合は形容詞（動詞）にも拡張し，名詞の場合はサ変名詞にも拡張する．但しサ変名詞だけの場合は拡張しない（例：テンプレートNo.5）．

表1　学習データ(新聞社説1年分)から得られた，厳しい制約によるテンプレート

No.	反復語	反復語間語列	後続語列
1	下一段動詞- サ変動詞- 五段動詞-	接続助詞バ-	形式名詞
2	上一段動詞- 五段動詞-	接続助詞バ-	(任意)
3	普通名詞- 指示名詞-	係助詞ハ-	格助詞デ
4	普通名詞- 指示名詞-	係助詞ハ- 格助詞ヲ-　格助詞ト-サ変動詞-接続助詞テ	
5	サ変名詞-	助動詞する-サ変名詞-サ変動詞-接続助詞テ-	助動詞する
6	普通名詞-	係助詞ハ	読点 格助詞ノ
7	普通名詞-	格助詞ガ-	格助詞デ-係助詞モ
8	形容詞-	形式名詞-係助詞ハ-	(任意)
9	普通名詞- 指示名詞-	係助詞モ-	(任意)
10	普通名詞-	助動詞ラシイ-	(任意)
11	五段動詞-	格助詞ニ-係助詞ハ-	(任意)
12	普通名詞-	格助詞ノ-形式名詞-格助詞ノ-	(任意)

表2　　表1から得られた，緩い制約によるテンプレート

No.	反復語	反復語間語列	後続語列
1,2	動詞- 形容詞-	接続助詞バ-	(任意)
4	名詞- サ変名詞-	係助詞ハ- 格助詞ヲ-　格助詞-サ変動詞-接続助詞	
5	サ変名詞-	助動詞する-サ変名詞-サ変動詞-接続助詞テ-	助動詞
5'	動詞- 形容詞-	サ変名詞-サ変動詞-接続助詞テ-	(任意)
3,6	名詞- サ変名詞-	係助詞ハ	記号系 格助詞 助動詞ダ
7	名詞- サ変名詞-	格助詞ガ-	格助詞-係助詞 助動詞ダ-係助詞
8	動詞- 形容詞-	形式名詞-係助詞ハ-	(任意)
9	名詞- サ変名詞-	係助詞モ-	(任意)
10	名詞- サ変名詞-	助動詞ラシイ-	(任意)
11	動詞- 形容詞-	格助詞ニ-係助詞ハ-	(任意)
12	名詞- サ変名詞-	格助詞ノ-形式名詞-格助詞ノ-	(任意)

［反復語間語列］

拡張しない．これは，反復語間語列が同語反復表現を特徴づける最も顕著な特徴であるので，拡張しないことが誤検出を避けるために得策と思われるためである．但し，反復語がサ変名詞に限られていて，それに続く反復語間語列の冒頭が「助動詞する」の場合は，反復語が動詞かつこの「助動詞する」が無い場合も同種の表現として成立可能と考えられるので，そのようなテンプレートを追加する（例：テンプレートNo.5'）．

［後続語列］

中品詞は大品詞に，小品詞は中品詞にそれぞれ拡張する．但し，格助詞デだけは，中品詞に拡張するのに加えて助動詞ダも加える（例：テンプレートNo.3，6，7）．これは格助詞デと助動詞ダに，用法の共通性が見られるためである．

3.2　テンプレートを用いた検出処理

本手法における検出処理の流れは図3の通りである．

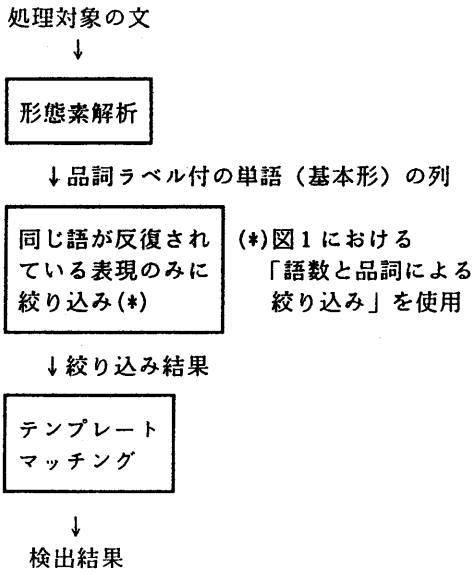

図3　検出処理の流れ

入力文に対して形態素解析を行い，同じ語が反復されている表現のみに絞り込んだ後，表1または表2のテンプレートにマッチする単語列を持つ文を検出する．テンプレートマッチングの前の絞り込みは，テンプレート作成の際に用いた「語数と品詞による絞り込み」をそのまま用いている．

3.3　インプリメント

検出システムは，Sparc Station 20 上で，jperl (Ver.4.019) を用いて記述した．処理に要する時間は，平均約0.35秒／文である．

4．評価

評価データとして，学習データとは異なる朝日新聞1993年社説1年分（21,843文）[6]を用いた．学習・評価データとして共に社説を用いたのは，新聞の一般記事では同語反復表現のような修辞的な表現が避けられるのに対し，論説やコラムでは比較的多用されると思われたからである．

一般に，検出機構の評価基準として，検出漏れ(第1種の誤り）の少なさを評価する呼出率(再現率)(recall factor)と，誤検出(第2種の誤り）の少なさを評価する適合率(precision factor)とがある．本稿ではこの両方について調べた．呼出率と適合率は，人手による検出結果との比較によって得ることになるが，そのために全評価データを人手で調べることは非現実的である．そこで，テンプレート作成の際に用いた「語数と品詞による絞り込み」を通過したものを，人手検出の対象とした．図4に，評価のための人手検出結果を得るための処理の流れを示す．つまり本稿における評価は，図3におけるテンプレートマッチング部の評価のことであり，その前の形態素解析および絞り込みでは検出漏れは発生していないと仮定している．

（1）厳しい制約によるテンプレートの評価

厳しい制約によるテンプレートで評価データを処理すると，表3より，呼出率（c/a）は63.9%，適合率（c/b）は100.0%となった．

検出漏れは，「言うだけ言う」「やむにやまれず」「やむにやまれぬ」「多ければ多いほど」「変われば変わるもの」「東は東」「

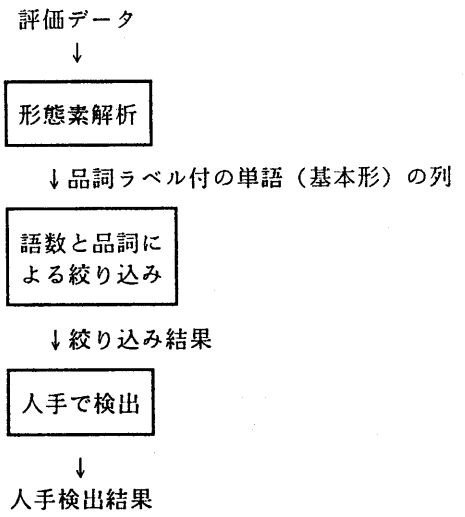

図4　評価のための人手検出結果を得る処理

西は西」「それならそれで」「出るべくして出た」「正直といえば正直だ」「変えるなら変えるで」「謝るだけ謝る」「民意は民意である」の１３個であった。一方、誤検出は皆無であった。

（２）緩い制約によるテンプレートの評価

緩い制約によるテンプレートで評価データを処理すると，表４より，呼出率（c/a）は80.6%，適合率（c/b）は93.5%となった．厳しい制約の場合と比較して検出漏れが６個減ったことで，呼出率については16.7%の改善が見られた．その６個は，「多ければ多いほど」「変われば変わるもの」「東は東」「西は西」「出るべくして出た」「民意は民意である」であった．緩い制約によっても検出できなかったのは残りの７つ，すなわち「言うだけ言う」「やむにやまれず」「やむにやまれぬ」「それならそれで」「正直といえば正直だ」「変えるなら変えるで」「謝るだけ謝る」であるが，これらは新しいテンプレートとして「～だけ～」「～に～ぬ」「～なら～」「～といえば～」の４つを追加登録するだけで検出できるようになる．以上より，今回の学習データから作成したテンプレートだけではまだ不十分であると言えるのと同時に，同語反復表現の検出に必要なテンプレートの種類はそれほど多くないことも示唆されているといえる．従って，学習データをより多くすれば，更に検出漏れの減少が期待できるものと思われる．

一方，誤検出は，「国民は国民に向かって」と「よければよい」の２個が発生した．前者は，「～は～」（テンプレートNo.3）の後続語列の制約（格助詞デ）を緩めた結果，すべての格助詞を受け入れてしまうようになったことによる誤りである．このように，制約の緩和は誤検出を増やす原因になってしまう問題がある．後者は，制約の緩和によってテン

表３　厳しい制約によるテンプレートを用いた場合の検出数

人手 a	機械 b	成功数 c(=a∩b)	検出漏れ a-c	誤検出 b-c
３６	２３	２３	１３	０

表４　緩い制約によるテンプレートを用いた場合の検出数

人手 a	機械 b	成功数 c(=a∩b)	検出漏れ a-c	誤検出 b-c
３６	３１	２９	７	２

プレートNo.1と2が統合されて，後続語列の形式名詞（例えば「多ければ多いほど」の「ほど」）が任意になったために拾い上げてしまった誤りである．つまり後続語列の中でテンプレートとして扱う範囲を明確にしなかったことに原因があったといえる．しかし今回の評価結果を見た限りでは，緩い制約を用いることで，検出漏れの減少による改善が，誤検出の増加による悪化を上回っていると言え，総合的には改善効果が期待できる．従って，緩め方をより工夫すれば，更に改善されるものと思われる．

5．本手法の課題

提案した手法には，以下のような課題が残されている．

（1）形態素解析において，表記上の揺れを吸収することを考慮していない．そのため，「私はワタシだ」のような，前反復語と後反復語の表記が異なる同語反復表現を漏らしてしまう問題がある．この問題の解決には，形態素解析器の改善が必要である．

（2）形態素解析の誤りによる検出漏れがある可能性が残っている．この問題の解決にも，形態素解析器の改善が必要である．

本手法のように，既製の形態素解析器を用いている限り，以上の２つの問題の解決の手だては無い．

（3）「表面的な処理では意味解析できない表現」という同語反復表現の定義が不明瞭である．そのため人手による検査では，同語反復表現であるか否かを，直観に頼っているという問題が残っている．しかし，同語反復表現の対象が増減しても，テンプレートマッチングの方法をとっている本手法では，テンプレートの追加／削除のみで対応できるので，提案した手法に本質的な改変を必要とはしない．

（4）現在の手法で検出できる同語反復表現は，反復語が１語の場合のみである．「美しい花は美しい花だ」のように，反復語が複数語の場合については，今後の課題とする．

（5）本稿で用いた評価データは，学習データと同じ新聞社の隣接した時期の社説であったため，より一般的な評価データを使った場合よりも良い結果が得られた可能性がある．他のコーパスを評価データとして試してみる必要がある．

（6）本手法の評価は，「語数と品詞による絞り込み」における検出漏れは皆無とみなして評価したが，本当に皆無かどうかを確認する必要がある．

6．まとめ

同語反復表現を計算機で検出する一手法を提案した．品詞の並びに関するテンプレートマッチングという簡易な手法により，80.6%の呼出率で検出できることを確認した．上記の課題を解決することが今後の目標である．

【謝辞】
形態素解析器Q_JPの使用を許諾して下さった（株）リコーと，開発者の亀田雅之氏に感謝致します．

【参考文献】
[1]滝澤修：「計算機によるトートロジーの意味理解」，言語処理学会第１回年次大会，B2-3，pp.161-164（Mar.1995）．
[2]新納他：「片方向の共起性による述語型関係表現の自動抽出」，自然言語処理，Vol.2，No.3，pp.73-86（1995）．
[3]新納他：「コーパスからの関係表現の自動抽出」，情報処理学会論文誌，Vol.35，No.11，pp.2258-2264（1994）．
[4]亀田雅之：「軽量・高速な日本語解析ツール『簡易日本語解析系Q_JP』」，言語処理学会第１回年次大会，P1-1，pp.349-352（Mar.1995）．
[5]朝日新聞フルテキスト記事データベース"CD-HIASK '94".
[6]朝日新聞フルテキスト記事データベース"CD-HIASK '93".