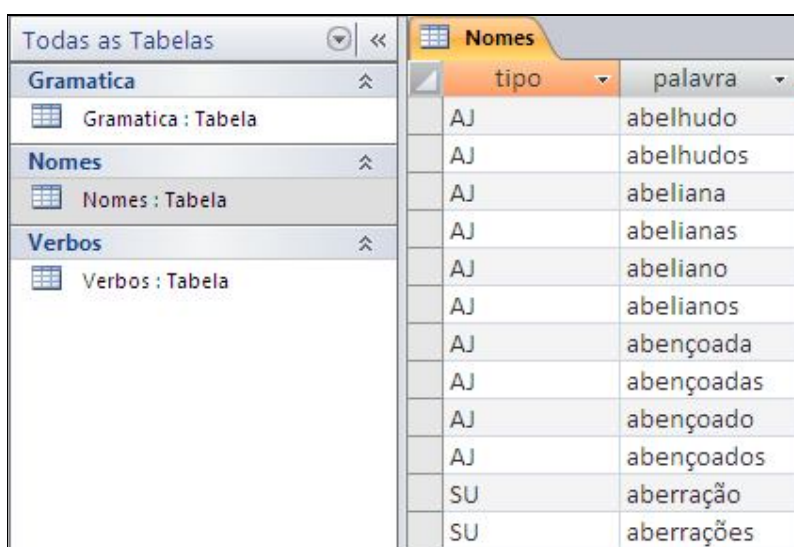


O primeiro desafio na construção desta ferramenta foi à elaboração de um léxico da língua Portuguesa completo o suficiente para permitir análises e conseqüente etiquetagem do texto.

A primeira etapa para a construção deste dicionário foi a adaptação de arquivos com o vocabulário utilizado pelo BR/ISPELL¹. Esta ferramenta foi desenvolvida para verificar a ortografia de projetos de código aberto. Através da adaptação destes arquivos foi possível a construção de um arquivo de dados (optou-se pelo uso do Access) com uma tabela de 41978 nomes e adjetivos.



tipo	palavra
AJ	abelhudo
AJ	abelhudos
AJ	abeliana
AJ	abelianas
AJ	abeliano
AJ	abelianos
AJ	abençoada
AJ	abençoadas
AJ	abençoado
AJ	abençoados
SU	aberração
SU	aberrações

FIGURA 3 – Dicionário do Ogma: Tabela de Nomes e Adjetivos.

Outro item necessário era a identificação dos verbos. Utilizando a ferramenta conjugue² e uma base de dados de 5000 verbos conseguiu-se reunir em outra tabela do banco de dados 292.720 verbos. Devido a regras de identificação de Sintagmas Nominais utilizadas nesta pesquisa também foram necessários identificar os verbos no particípio, estes verbos receberam uma identificação diferenciada “VP” no lugar de “VB”. A próxima ilustração demonstra parte da tabela de verbos, e o verbo “abafado” da figura mostra como a diferenciação foi feita.

¹ Para mais informações: <http://www.ime.usp.br/~ueda/br.ispell>

² O conjugue é um script *awk* capaz de conjugar verbos da língua portuguesa, a partir de um banco de paradigmas.

Todas as Tabelas		Verbos	
Gramatica		palavra	tipo
Gramatica : Tabela		abafa	VB
		abafa	VB
		abafado	VP
		abafai	VB
		abafais	VB
		abafam	VB
		abafamos	VB
		abafamos	VB
		abafando	VB
		abafar	VB

FIGURA 4 – Dicionário do Ogma: Tabela de Verbos.

Finalmente através de um processo manual de digitação, tendo como base a gramática de Tufano (1990) conseguiram-se reunir 475 palavras de diversas classes gramaticais. Estas palavras são as mesmas que são utilizadas para compor a lista de *stopwords* (Anexo II).

Todas as Tabelas		Gramatica	
Gramatica		Palavra	Tipo
Gramatica : Tabela		agora	AV
		ah	IT
		ai	IT
		ainda	AV
		além	AV
		além+de	PR
		algo	PI
		alguém	PI
		algum	PI
		ali	AV
		alve	IT
		amanhã	AV
		amém	IT
		ânimo	IT

FIGURA 5 – Dicionário do Ogma: Tabela Gramatical.

Este processo de identificação das classes gramaticais será útil na etiquetagem do texto analisado conforme o próximo quadro demonstra a associação das etiquetas utilizadas no Ogma com as etiquetas propostas no modelo ED-CER.

ETIQUETA OGMA	ETIQUETA ED-CER	CLASSE GRAMATICAL
AD	AR	Artigo definido

AI	AR	Artigo indefinido
AJ	AJ	Adjetivo
AV	AV	Advérbio
CJ	CO	Conjunção (Aditiva, adversativa, alternativa)
IT	*	Interjeição
NC	NU	Números cardinais
NM	NU	Números multiplicativos
NO	NU	Números ordinais
NP	NP	Nome próprio
NR	NU	Número romano
PS	PS	Pronome possessivo
PD	PD	Pronome demonstrativo
PI	PI	Pronome indefinido
PL	LG	Pronome relativo
PN	PN	Pontuação (exceto vírgula)
PP	PP	Pronome pessoal
PR	PR	Preposições
SU	SU	Substantivo
VB	VB	Verbos
VG	CO	Vírgulas
VP	AJ	Verbos no Particípio

QUADRO 1 – Etiquetas utilizadas no Ogma e no ED-CER.

O texto deve ser tratado previamente, transformando expressões para melhorar o processo de etiquetagem. O seguinte conjunto é utilizado no Ogma:

do que → que
 ao → a o
 aos → a os
 pela → por a
 pelas → por a
 pelos → por o
 pelo → por e
 neste → em este
 nisto → em isto
 nuns → em uns
 numas → em umas
 num → em um
 numa → em uma
 duns → de uns
 dumas → de umas
 dum → de um
 duma → de uma
 nos → em os
 dos → de os
 do → de o
 das → de as
 da → de a
 nas → em as
 no → em o
 na → em a
 ás → aos as
 a fim de → a+fim+de

a que → a+que
 a qual → a+qual
 a respeito de → a+respeito+de
 abaixo de → abaixo+de
 acerca de → acerca+de
 acima de → acima+de
 além de → além+de
 antes de → antes+de
 ao invés de → ao+invés+de
 ao redor de → ao+redor+de
 apesar de → apesar+de
 as quais → as+quais
 até a → até+a
 bem como → bem+como
 como também →
 como+ também
 como um → como+um
 de acordo com →
 de+acordo+com
 debaixo de → debaixo+de
 defronte de → defronte+de
 dentreo de → dentreo+de
 depois de → depois+de
 diante de → diante+de
 e não → e+não
 em cima de → em+cima+de
 em face de → em+face+de

em frente a → em+frente+a
 em frente de → em+frente+de
 em lugar de → em+lugar+de
 em vez de → em+vez+de
 mas ainda → mas+ainda
 mas também → mas+também
 não obstante → não+obstante
 não obstante → não+obstante
 não sí → não+sí
 não só → não+só
 no caso de → no+caso+de
 no entanto → no+entanto
 o qual → o+qual
 o que → o+que
 os quais → os+quais
 para com → para+com
 perto de → perto+de
 por conseguinte →
 por+conseguinte
 por isso → por+isso
 por trás de → por+trás+de

O próximo desafio depois de elaborado o dicionário foi o tratamento das ambigüidades. Por exemplo, a palavra “mato” pode ser Verbo ou Substantivo dependendo do contexto e da posição.

Exemplos:

*Eu **mato** o rato*

e

*O **mato** estava grande*

Para contornar esta dificuldade o Ogma forma uma lista com todas as combinações encontradas e submete frase a frase às regras para extração dos SNs.

Texto etiquetado pelo Ogma:

- 1) *Eu/PP mato/VBSU o/AD rato/AJSU*
- 2) *O/AD mato/VBSU estava/VB grande/AJ*

Então o OGMA submete às regras de extração definidas, duas versões de cada frase. No caso do primeiro exemplo são enviadas duas possibilidades:

Eu/PP mato/VB o/AD rato/AJ

Eu/PP mato/SU o/AD rato/AJ

Os Sintagmas Nominais encontrados entram em uma lista geral de Sintagmas Nominais da frase, eliminado os duplicados. Este tratamento possibilitou resolver o problema da ambigüidade de forma bem eficiente.

Para extrair os SNs, o seguinte conjunto de regras foi utilizado, regra por regra na ordem abaixo:

AR ← AD	
AR ← AI	
AJ ← VP	
NU ← NR	
NU ← NC	
CO ← VG	
	AV ←AV ad
	MD ←AV MD
	MD ←MD co MD
	NS ←NS MD

CO ← CJ	NS ←MD NS
de ← AR	NS ←NS pr NS
de ← PD	NS ←NS pr de NS
de ← PI	NS ←NS co NS
qu ← AJ	NS ←NS co de NS
qu ← NU	NS ←AV NS
qu ← PS	SN ←de SN
ad ← AV	
co ← CO	
pr ← PR	
re ← SU	
de ← PP	
re ← NP	
NS ←re	
MD ←qu	
SN ←NS	
AV ←ad	

QUADRO 2 – Regras de extração de SN do método OGMA.

Uma regra foi modificada em relação às regras do ED-CER: nova regra: **de ← PP** substituindo a regra do ED-CER **re ← PP**. A modificação desta regra se deu para melhorar a extração de SN com base em diversos testes executados.

Como o OGMA é uma ferramenta projetada para auxiliar na execução de todo o experimento proposto na metodologia, esta ferramenta deveria ser capaz de:

- Extrair os Sintagmas Nominais.
- Atribuir pesos aos Sintagmas Nominais extraídos de acordo com a frequência que aparecem no texto.
- Atribuir pesos aos Sintagmas Nominais extraídos de acordo com a frequência que aparecem no texto e dentro de outros Sintagmas Nominais.
- Identificar a classe do Sintagma Nominal (CSN) extraído de acordo com a metodologia proposta por SOUZA (2005) e explicada no item 2.3.4 deste trabalho.
- Calcular a pontuação de cada Sintagma Nominal extraído (relevância como descritor) utilizando a mesma metodologia.
- Extrair termos e atribuir pesos de acordo com sua frequência no texto.
- Extrair termos, exceto os constantes na lista de *Stopwords*, e atribuir pesos de acordo com sua frequência no texto.

- h) Calcular a similaridade entre duas listas de termos (extraídas do documento) utilizando o coseno.

Criaram-se diversos comandos, cada um realizando uma função específica. O primeiro recurso trabalhado foi o de etiquetagem (anotação) do texto. Para utilizá-lo no Ogma, devem-se utilizar três parâmetros: o primeiro o comando "E", o segundo o nome do arquivo origem em formato texto e o terceiro parâmetro que seria o arquivo de saída na qual será armazenado o texto etiquetado.

Sintaxe:

"ogma e textooriginal.txt textoetiquetado.txt"

Em seguida criou-se a opção para extração de Sintagmas Nominais, opção "S". Esta opção analisa o arquivo de entrada, que já deverá estar etiquetado e faz a aplicação das regras de extração. Os Sintagmas nominais são salvos então na ordem em que aparecem em outro arquivo. Para se utilizar esta função deve-se fornecer como primeiro parâmetro o "S" em seguida o nome do arquivo contendo o texto já etiquetado e o terceiro parâmetro o arquivo de saída.

Sintaxe:

"ogma s textoetiquetado.txt relacaosn.txt"

Também se criou uma opção para visualização rápida dos sintagmas nominais sem necessitar de passar pelas duas etapas anteriores. Esta opção "X" recebe apenas um parâmetro relativo ao texto que se pretende analisar.

Sintaxe:

"ogma x textooriginal.txt"

A próxima etapa de implementação foi às opções que permitiriam gerar as tabelas de termo e peso para cada um dos métodos propostos.

A primeira opção criada foi a que gera uma tabela contendo na primeira coluna a lista de todos os termos utilizados, na segunda coluna o número de vezes na qual o termo aparece em todo o texto, e na terceira coluna a porcentagem que aquele termo responde na composição de todo o documento. Para utilizar esta opção, "TT", deve-se fornecer como primeiro parâmetro o texto a ser analisado, e como segundo parâmetro o arquivo de saída.

Sintaxe:

"ogma tt textooriginal.txt tabtermos.txt"

Seguindo a mesma linha criou-se a opção para geração da tabela ignorando as palavras que aparecem na lista de *stopwords*. A relação completa das palavras que compõem esta lista encontra-se no Anexo II deste trabalho. Para gerar esta tabela, utiliza-se a opção "TTS", com dois parâmetros: o primeiro o arquivo texto de entrada e o segundo o arquivo de saída na qual será armazenada a tabela. O arquivo de saída possui também 3 colunas e segue a especificação da tabela gerada pela opção de termos, "TT".

Sintaxe:

"ogma tts textooriginal.txt tabtermos.txt"

A primeira opção de geração de tabelas relativas a Sintagmas Nominais é a "TS". Nesta opção também é gerada uma tabela com três colunas: a primeira contém uma lista de Sintagmas Nominais, a segunda o número de vezes que aquele Sintagma Nominal aparece no texto, e a terceira o cálculo em relação a todo o documento. A opção trabalha com dois parâmetros: o primeiro deverá ser a relação de sintagmas nominais, um em cada linha, e o segundo o arquivo de saída. A relação de sintagmas nominais é gerada pela opção "S" como visto anteriormente.

Sintaxe:

"ogma ts relacaodesn.txt tabsn.txt"

Esta opção, entretanto considera apenas Sintagmas Nominais únicos. Se dois Sintagmas Nominais são localizados, por exemplo, "Gestão" em um parágrafo e "Gestão do conhecimento" em outro a opção "TS" não considerará que o primeiro SN "Gestão" apareceu duas vezes. Optou-se por criar uma segunda opção "TC" que contabilizasse também os sintagmas nominais dentro dos outros sintagmas nominais encontrados. Esta opção utiliza os mesmo parâmetros da opção anterior a "TS".

Sintaxe:

"ogma tc relacaodesn.txt tabsn.txt"

A próxima tabela seria a opção "TR" que gera uma tabela com os Sintagmas Nominais pontuados de acordo com a metodologia proposta por SOUZA (2005). Para cumprir esta função, foi preciso etiquetar novamente a relação de sintagmas nominais, fornecida como parâmetro, para descobrir a classe de sintagma nominal (CSN), item necessário para o cálculo da pontuação. Esta etiquetagem é realizada internamente

pelo OGMA para que os parâmetros permanecessem os mesmo das opções de geração de tabelas de sintagmas nominais. A tabela segue o mesmo formato das anteriores com o acréscimo de uma quarta coluna onde é salva a classificação (CSN) encontrada para o sintagma nominal.

Sintaxe:

"ogma tr relacaodesn.txt tabsn.txt"

O processo de extração de SN realizada pelo método ED-CER resulta em uma lista de SNs na sua forma máxima. Nesta pesquisa utilizou-se também a opção de SN Aninhados. O SN Aninhado considera também como descritor os núcleos que compõe o SN máximo.

Por exemplo:

"A gestão do conhecimento nas organizações nacionais" corresponde a um SN máximo.

Entretanto existem três SN Aninhados:

- 1) "A Gestão"
- 2) "conhecimento"
- 3) "as organizações nacionais."

Realizou-se então uma adaptação no método "TR", descrito acima, para considerar não só o Sintagma Nominal Máximo, mas também todos os aninhados. Duas opções foram criadas: a "TCA" que realiza a extração dos SN mais os aninhados e a "TRA" que além desta extração os pontua na metodologia de Souza (2005).

Sintaxe:

"ogma tra relacaodesn.txt tabsn.txt" ou "ogma tca relacaodesn.txt tabsn.txt"

A próxima implementação seria relativa ao cálculo da similaridade (discutidas no item 2.4.3). Este cálculo faz a aplicação da fórmula do coseno em duas tabelas, cada uma relativa a um arquivo, e retorna valores próximos a 1 à medida que os dois documentos comparados são semelhantes. Quanto mais próximo de 1 mais semelhante é um documento do outro. O resultado 1 significa que os documentos são iguais.

Sintaxe:

"ogma i tabela1.txt tabela2.txt"

Para facilitar a utilização do ogma, sem a necessidade de ter que se passar por vários comandos e etapas para calcular a similaridade, criaram-se três opções

adicionais "IT", "IC" e "IR" que calculam respectivamente a similaridade entre dois textos utilizando os métodos: por termos, por sintagmas nominais e por sintagmas nominais pontuados.

No ANEXO encontram-se exemplos de utilização, contendo arquivos de entrada e arquivos de saída, de cada uma das opções do OGMA descritas anteriormente.

A próxima figura demonstra a utilização do OGMA para o cálculo da similaridade entre dois documentos.

```
C:\WINDOWS\system32\cmd.exe

OGMA v0.7
Ferramenta para análise de texto.
=====
= Desenvolvido por Luiz Cláudio Maia =
= luizmaia@luizmaia.com.br =
=====
= Orientação Renato Rocha Souza, EGI/UFMG =
=====
Analisando palavras do arquivo in-teste.txt
Escrevendo tabela no arquivo temp1.$$$
Número de palavras analisadas 43
Número de termos analisadas 39
Analisando palavras do arquivo in-dgz.txt
Escrevendo tabela no arquivo temp2.$$$
Número de palavras analisadas 3132
Número de termos analisadas 1101
Comparando o arquivo temp1.$$$ com temp2.$$$
Similaridade (cos): 0,264153
=====
C:\Projetos\Net\Ogma\Ogma\bin\Debug>
```

FIGURA 6 – Resultado do cálculo da similaridade entre dois documentos pelo Ogma.

Comparações com outras ferramentas existentes para extração de sintagmas nominais com o OGMA foram realizadas. Quando aplicamos o texto abaixo, nos métodos ED-CER, OGMA e VISL obtêm a extração de SN iguais para os métodos ED-CER (manual) e OGMA (automático), comprovando que a automação foi eficaz.

ANEXO - Exemplos de utilização do OGMA

e – Etiquetar texto

exemplo: ogma e texto.txt textoetiquetado.txt

ENTRADA	SAÍDA
O novo cálculo das aposentadorias resulta em valores menores do que os atuais para quem perde o benefício com menos tempo de contribuição e idade.	o/AD novo/AJ cálculo/SU de/PR as/AD aposentadorias/SU resulta/VB em/PR valores/SU menores/AJSU que/CJPL os/ADPR atuais/VBAJSU para/PR quem/PL perde/VB o/AD benefício/SU com/PR menos/AV tempo/SU de/PR contribuição/SU e/CJ idade/SU ./PN

s - Extraí os Sintagmas Nominais e grava em um arquivo

exemplo: ogma s textoetiquetado.txt relacaosn.txt

ENTRADA	SAÍDA
o/AD novo/AJ cálculo/SU de/PR as/AD aposentadorias/SU resulta/VB em/PR valores/SU menores/AJSU que/CJPL os/ADPR atuais/VBAJSU para/PR quem/PL perde/VB o/AD benefício/SU com/PR menos/AV tempo/SU de/PR contribuição/SU e/CJ idade/SU ./PN	- o novo cálculo das aposentadorias - o benefício com menos tempo de contribuição e idade - valores menores que os atuais

x – Mostra os Sintagmas Nominais do arquivo

ex: ogma x texto.txt

ENTRADA	SAÍDA
O novo cálculo das aposentadorias resulta em valores menores do que os atuais para quem perde o benefício com menos tempo de contribuição e idade.	- o novo cálculo das aposentadorias - o benefício com menos tempo de contribuição e idade - valores menores que os atuais

tt - gera tabela de termos com numero de vezes que aparecem no texto

ex: ogma tt texto.txt tabtermos.txt

ENTRADA	SAÍDA
O novo cálculo das aposentadorias resulta em valores menores do que os atuais para quem perde o benefício com menos tempo de contribuição e idade.	novo/1/0,055556 cálculo/1/0,055556 das/1/0,055556 aposentadorias/1/0,055556 resulta/1/0,055556 valores/1/0,055556 menores/1/0,055556 que/1/0,055556 atuais/1/0,055556 para/1/0,055556 quem/1/0,055556 perde/1/0,055556 benefício/1/0,055556 com/1/0,055556 menos/1/0,055556 tempo/1/0,055556 contribuição/1/0,055556 idade/1/0,055556

tts - gera tabela de termos com número de vezes que aparecem no texto (filtra stopwords)

ex: ogma tts texto.txt tabtermos.txt

ENTRADA	SAÍDA
O novo cálculo das aposentadorias resulta em valores menores do que os atuais para quem perde o benefício com menos tempo de contribuição e idade.	novo/1/0,076923 cálculo/1/0,076923 das/1/0,076923 aposentadorias/1/0,076923 resulta/1/0,076923 valores/1/0,076923 menores/1/0,076923 atuais/1/0,076923 perde/1/0,076923 benefício/1/0,076923 tempo/1/0,076923 contribuição/1/0,076923 idade/1/0,076923

ts - gera tabela de sn etiquetados com n. de vezes que aparecem

ex: ogma ts relacaosn.txt tabsnf.txt

ENTRADA	SAÍDA
O novo cálculo das aposentadorias resulta em valores menores do que os atuais para quem perde o benefício com menos tempo de contribuição e idade.	o novo cálculo das aposentadorias/1/0,333333 o benefício com menos tempo de contribuição e idade/1/0,333333 valores menores que os atuais/1/0,333333

tc - gera tabela de SN etiquetados com número de vezes que aparecem em todo o texto

ex: ogma tc relacaosn.txt tabsnf.txt

ENTRADA	SAÍDA
O novo cálculo das aposentadorias resulta em valores menores do que os atuais para quem perde o benefício com menos tempo de contribuição e idade.	o novo cálculo das aposentadorias/1/0,333333 o benefício com menos tempo de contribuição e idade/1/0,333333 valores menores que os atuais/1/0,333333

tr - gera tabela de SN pontuados

ex: ogma tr relacaosn.txt tabsnf.txt

ENTRADA	SAÍDA
- o novo cálculo das aposentadorias - o benefício com menos tempo de contribuição e idade - valores menores que os atuais	o novo cálculo das aposentadorias/1/1,100000/2 o benefício com menos tempo de contribuição e idade/1/1,400000/3 valores menores que os atuais/1/1,100000/2

i - calcula a similaridade entre duas tabelas de termos

ex: ogma i tabela1.txt tabela2.txt

it - calcula a similaridade entre dois textos comparando por termos

ex: ogma i texto1.txt texto2.txt

ir - calcula a similaridade entre dois textos comparando com SN pontuados

ex: ogma i texto1.txt texto2.txt

ic - calcula a similaridade entre dois textos comparando com os SN

ex: ogma i texto1.txt texto2.txt

