ENEPEX
ENCONTRO DE ENSINO, PESQUISA E EXTENSÃO
8° ENEPE UFGD • 5° EPEX UEMS

DESENVOLVIMENTO DE PROTÓTIPO PARA DETECÇÃO DE SENTENÇAS DE CORPUS PARA POPULAR ONTOLOGIA NÚCLEO

Joinvile Batista Junior[1]; Paulo Edson Misokami Oliveira[2];
UFGD-FACET, C. Postal 533, 79804-970 Dourados-MS, E-mail: joinvile@ufgd.edu.br
[1] Professor do curso de Sistemas de Informação. [2] Aluno voluntário de Iniciação Científica.

RESUMO

Ontologias, utilizadas para representar conhecimento em um dado domínio, são construídas automaticamente a partir de relações geradas por extratores de informação, a partir de um corpus formado por um ou vários arquivos contendo informação textual. A internet corresponde a um gigantesco corpus, com informações consistentes convivendo com informações conflitantes e desatualizadas, associadas a fatos ou a distintos pontos de vista. A inserção de relações incorretas, ou contraditórias em uma ontologia, invalida a sua utilização para consultas e inferências. A construção automática de uma ontologia, a partir de uma fonte textual fidedigna, favorece sua validação posterior. O corpus escolhido, como fonte fidedigna, é um livro de referência para o ensino de nutrição básica. No entanto, a conversão do formato PDF deste corpus para o formato TXT, para realizar a extração de sentenças, resulta em um grande desafio para a detecção de sentenças, em função da mistura de texto associado a: títulos, parágrafos de seções, anotações, definições, textos de figuras, textos de tabelas, e referências. A utilização de detectores de sentenças disponíveis, a partir do formato TXT do referido corpus, resulta em um grande número de sentenças incorretas associada a uma grande perda de informação, o que inviabiliza a tarefa de extração de informação, que é essencial para a construção automática de uma ontologia. O objetivo deste trabalho é o desenvolvimento de um protótipo para extrair corretamente as sentenças dos capítulos do corpus. A metodologia utilizada incorpora: a identificação dos problemas para isolar o texto das sentenças das seções dos capítulos, dos demais tipos de ocorrências nas páginas do corpus; a solução proposta para os problemas apontados; e a prototipagem da solução proposta para extração dos trechos de texto e de suas sentenças. Foram extraídas sentenças corretamente dos primeiros quatro capítulos do corpus.

Palavras-chave: processamento de linguagem natural, automação de ontologias, detecção de sentenças.

INTRODUÇÃO

O experimento utilizado neste trabalho é baseado em textos da língua inglesa. Por este motivo, todos os exemplos apresentados no decorrer deste texto, são extraídos de sentenças da língua inglesa.

Ontologias representam conceitos (entidades), de um dado domínio geral ou específico, interligados por relações. A construção automática de ontologias depende da extração de relações de fontes textuais (SARAWAGI, 2008, p. 261-377). Ontologias construídas automaticamente a partir de fontes semi-estruradas, como a Wikipedia, tem reportados alta acurácia: *YAGO* (SUCHANEK; KASNECI; WEIKUM, 2008, p. 203-217) e *KOG Ontology* (WU; WELD, 2008, p. 635-644). Fontes semi-estruturadas seguem um determinado padrão, que facilita a automação da extração de informação dos textos. A extensão automática de ontologias implica na inclusão de novas relações, que necessitam ser validadas de forma a não incluir relações contraditórias. No sistema *SOFIE* (SUCHANEK; KASNECI; WEIKUM, 2009, p. 631-640), foi relatada alta acurácia na expansão da ontologia *YAGO*, construída automaticamente a partir de uma fonte semi-estruturada, a partir de informação textual não estruturada. No entanto, a construção automática de ontologias, a partir de fontes não estruturadas, permanece um grande desafio para a pesquisa na área de processamento de linguagem natural.

Extratores de informação como, por exemplo, *ReVerb* (FADER; SODERLAND; ETZIONI, 2011, p. 1535–1545) e R2A2 (ETZIONI et al., 2011, p. 3-10), extraem relações binárias a partir de sentenças. Na sentença “John gives out lots of candy on Halloween to the kids on his block.”, a relação binary é: *John* (entidade 1) – *gives out* (relação) – *lots of candy* (entidade 2). Rotuladores de papéis semânticos como, por exemplo, *UIUC-SRL* (PUNYAKANOK; ROTH; YIH, 2008, p. 1-30) e *Lund-SRL* (JOHANSSON; NUGUES, 2008, p. 393-400), assinalam os papéis semânticos que correspondem a relações n-árias a partir de sentenças. Para mesma sentença, os papéis semânticos são: *John* (distribuidor) – *gives out* (relação) – *lots of candy* (coisa distribuída) – *on Halloween* (modificador temporal) – *to the kids* (distribuído para) – *on his block* (modificador de localização).

A extração de relações de uma fonte não estruturada tem como entrada as sentenças extraídas do corpus selecionado. Corpus gerados manualmente são muito utilizados na pesquisa de processamento de linguagem natural. Extratores de informação têm se beneficiado do corpus *Penn Tree Bank* (KINGSBURY; PALMER, 2002, p. 1989-1993), gerado manualmente a partir de notícias de jornais. Rotuladores de papéis semânticos tem utilizado o corpus *PropBank* (KINGSBURY; PALMER; GILDEA, 2004, p. 1-33), que foi gerado a partir da rotulação manual de papéis semânticos do corpus *Penn Tree Bank*.

A construção automática de uma ontologia, em dado domínio, requer a utilização de um corpus que forneça um significativo conjunto de sentenças vinculadas ao domínio de interesse. A internet representa um gigantesco corpus, com informações consistentes convivendo com informações

conflitantes e desatualizadas, associadas a fatos ou a distintos pontos de vista. No entanto, as sentenças do corpus para a criação de uma ontologia devem abordar um determinado domínio de aplicação. Adicionalmente, a validação de sentenças extraídas da internet é uma tarefa bastante complexa.

A estratégia adotada para a construção e extensão automática de uma ontologia, no projeto de pesquisa ao qual se vincula este trabalho, é a construção automática de uma ontologia núcleo a partir de uma fonte textual fidedigna, para posterior extensão automática a partir de textos capturados na internet. O corpus escolhido foi a décima segunda edição de um livro de referência para o ensino de nutrição básica (WHITNEY; ROLFES, 2011; 1007 p.).

Para um corpus composto de um conjunto de sentenças, a tarefa de detecção de sentenças é relativamente simples. O *toolkit Apache OpenNLP* possui um detector de sentenças bastante utilizado (APACHE, 2010). No entanto, a conversão do formato PDF do corpus escolhido para o formato TXT, para realizar a extração de sentenças, resulta em um grande desafio para a detecção de sentenças, em função da mistura de texto associado a: títulos, parágrafos de seções, anotações, definições, textos de figuras, textos de tabelas, e referências. A utilização de detectores de sentenças disponíveis, a partir do formato TXT do referido corpus, resulta em um grande número de sentenças incorretas associada a uma grande perda de informação, o que inviabiliza a tarefa de extração de informação, que é essencial para a construção automática de uma ontologia. A extração incorreta de sentenças pode resultar em sentenças compostas a partir de partes de sentenças distintas da fonte original, comprometendo irremediavelmente a extração de relações e, como consequência, a tarefa de construção automática de uma ontologia.

Neste trabalho, foi desenvolvido um protótipo para suportar a detecção correta de sentenças, dos capítulos do corpus, de forma viabilizar as tarefas posteriores de extração de informações e de construção automática da ontologia núcleo. Como entrada para este protótipo, é utilizado o conversor do formato PDF para TXT, *Foxit Reader* (FOXIT, 2013), que organiza melhor a informação do ponto de vista espacial, criando duas colunas de informação, se comparado com o conversor tradicional suportado pelo *Adobe Reader* (ADOBE, 2014).

1. DESENVOLVIMENTO

A metologia utilizada, para o desenvolvimento deste trabalho, abrange três etapas: (a) a identificação dos problemas para isolar o texto das sentenças das seções dos capítulos, dos demais tipos de ocorrências nas páginas do corpus; (b) a solução proposta resolver os problemas apontados; e (c) a prototipagem da solução proposta para extração dos trechos de texto e de suas sentenças.

Devido à natureza deste trabalho, é imprescindível a ilustração de trechos do corpus, nos formatos PDF e TXT, para caracterizar, de forma mais legível, cada um dos problemas encontrados, e para explicar a respectiva solução. Infelizmente, em função do espaço ocupado pelas figuras no texto

deste trabalho, foi adotada uma estratégia baseada em duas premissas, para a seleção das figuras utilizadas para ilustrar este trabalho. A primeira premissa é a escolha somente das figuras mais relevantes para o entendimento do trabalho, de forma que não será possível ilustrar todos os problemas e soluções abordados neste trabalho. A segunda premissa é a representação somente dos trechos mais representativos de uma página, necessário para caracterizar o problema ou a solução em questão, dado que a representação de toda a página é inviável.

1.1 IDENTIFICAÇÃO DOS PROBLEMAS PARA ISOLAR AS SENTENÇAS

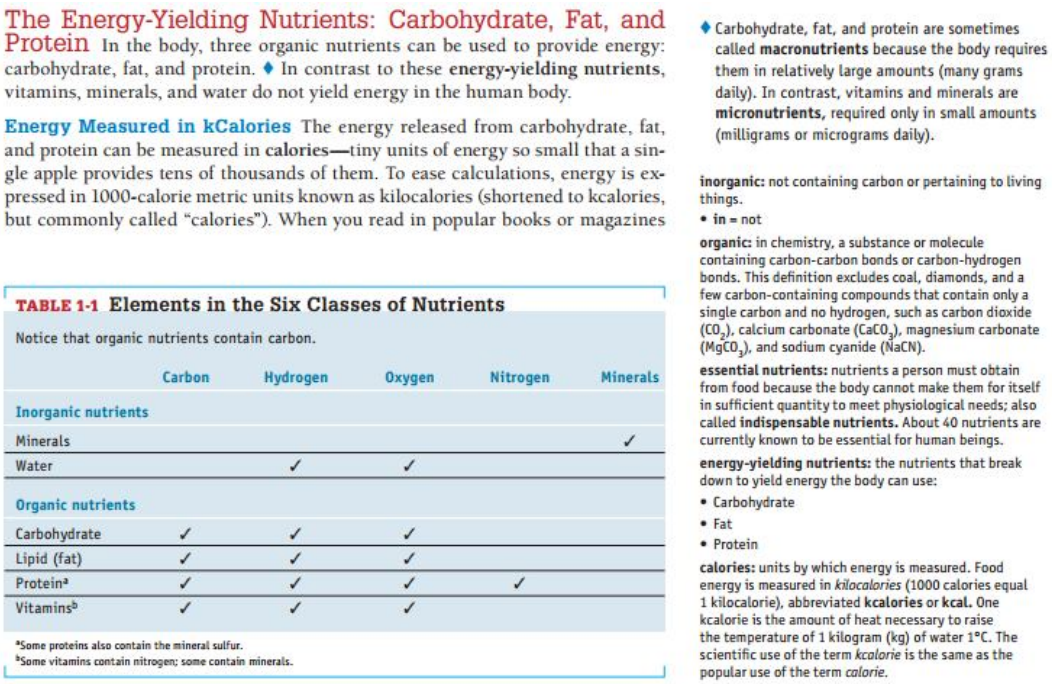
A primeira etapa do trabalho é a identificação dos vários tipos de ocorrências de informação textual, gráfica (figuras isoladas) e mista (tabelas e figuras associadas a textos), que aparecem no formato TXT, gerado a partir do corpus no formato PDF, e que dificultam a separação dos trechos de texto que contém as sentenças de interesse para extração. À medida que os problemas são identificados, é atribuído um rótulo para caracterizar adequadamente o problema e para conectar a sua identificação com a solução proposta para resolvê-lo, descrita na seção seguinte.

No mais alto nível, o corpus utilizado é composto de: capítulos, highlights, apêndices e glossários. *Highlights* são textos complementares de cada capítulo, sempre ao final do respectivo capítulo. Cada capítulo é composto de seções e de definições.

A tarefa de automatizar a extração das sentenças do corpus é complexa em função da grande diversidade de situações que precisam ser tratadas para isolar adequadamente as sentenças de interesse. Este trabalho é o primeiro passo na direção do objetivo de extrair todas as sentenças do corpus, que possam ser posteriormente utilizadas para a extração de relações e construção de uma ontologia no domínio de nutrição. O foco deste trabalho inicial é a extração das sentenças dos capítulos, que representam a informação mais relevante do ponto de vista de um protótipo inicial de extração de sentenças do corpus.

Na figura 1 é ilustrado um trecho de uma página típica no formato PDF. Observa-se, do ponto de vista da distribuição espacial dos conteúdos no trecho ilustrado, que a página é composta por duas colunas, com larguras distintas. Neste trecho, somente os parágrafos do canto superior esquerdo da coluna esquerda são utilizados para a extração de sentenças. Estes parágrafos precisam ser isolados da tabela que os sucede na mesma coluna e das definições que aparecem na coluna direita. Definições são caracterizadas por: (a) parágrafos explicativos, referenciados no texto das sentenças pelo símbolo "♦"; ou por (b) termos, compostos de uma ou mais palavras, cujo sentido é caracterizado por uma sentença ou um trecho de texto, separados pelo caracter ":". Nesta página, a coluna, que contém as seções do capítulo, ocupa um espaço maior na página, em relação às colunas que contém as definições. Estas

colunas são alternadas a cada página. Na próxima página, as definições são posicionadas na coluna esquerda, que passa ocupar o menor espaço na página, e as seções são posicionadas na coluna direita.



The Energy-Yielding Nutrients: Carbohydrate, Fat, and Protein In the body, three organic nutrients can be used to provide energy: carbohydrate, fat, and protein. ♦ In contrast to these **energy-yielding nutrients**, vitamins, minerals, and water do not yield energy in the human body.

Energy Measured in kCalories The energy released from carbohydrate, fat, and protein can be measured in **calories**—tiny units of energy so small that a single apple provides tens of thousands of them. To ease calculations, energy is expressed in 1000-calorie metric units known as kilocalories (shortened to kcalories, but commonly called "calories"). When you read in popular books or magazines

TABLE 1-1 Elements in the Six Classes of Nutrients

Notice that organic nutrients contain carbon.

	Carbon	Hydrogen	Oxygen	Nitrogen	Minerals
Inorganic nutrients					
Minerals					✓
Water		✓	✓		
Organic nutrients					
Carbohydrate	✓	✓	✓		
Lipid (fat)	✓	✓	✓		
Protein[a]	✓	✓	✓	✓	
Vitamins[b]	✓	✓	✓		

[a]Some proteins also contain the mineral sulfur.
[b]Some vitamins contain nitrogen; some contain minerals.

♦ Carbohydrate, fat, and protein are sometimes called **macronutrients** because the body requires them in relatively large amounts (many grams daily). In contrast, vitamins and minerals are **micronutrients,** required only in small amounts (milligrams or micrograms daily).

inorganic: not containing carbon or pertaining to living things.
- **in** = not

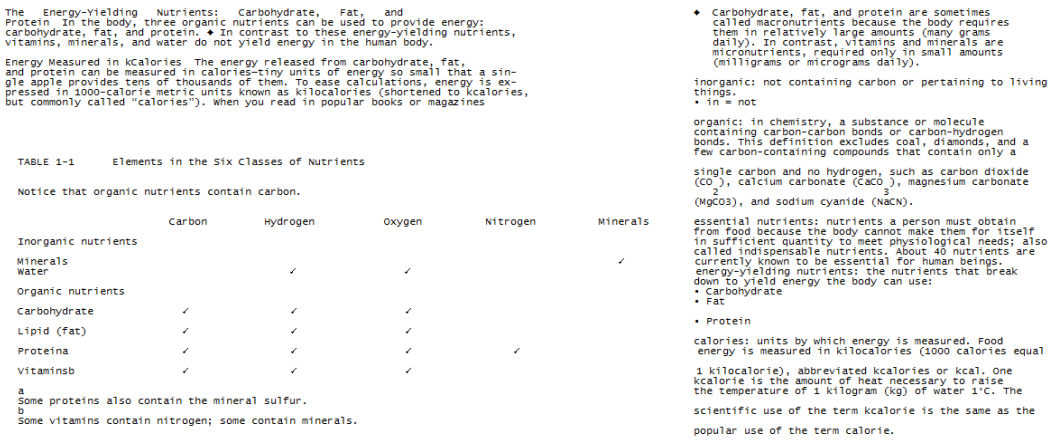
organic: in chemistry, a substance or molecule containing carbon-carbon bonds or carbon-hydrogen bonds. This definition excludes coal, diamonds, and a few carbon-containing compounds that contain only a single carbon and no hydrogen, such as carbon dioxide (CO_2), calcium carbonate ($CaCO_3$), magnesium carbonate ($MgCO_3$), and sodium cyanide (NaCN).

essential nutrients: nutrients a person must obtain from food because the body cannot make them for itself in sufficient quantity to meet physiological needs; also called **indispensable nutrients.** About 40 nutrients are currently known to be essential for human beings.

energy-yielding nutrients: the nutrients that break down to yield energy the body can use:
- Carbohydrate
- Fat
- Protein

calories: units by which energy is measured. Food energy is measured in *kilocalories* (1000 calories equal 1 kilocalorie), abbreviated **kcalories** or **kcal.** One kcalorie is the amount of heat necessary to raise the temperature of 1 kilogram (kg) of water 1°C. The scientific use of the term *kcalorie* is the same as the popular use of the term *calorie.*

FIGURA 1: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 29)
com Página Típica com duas Colunas



The Energy-Yielding Nutrients: Carbohydrate, Fat, and Protein In the body, three organic nutrients can be used to provide energy: carbohydrate, fat, and protein. ♦ In contrast to these energy-yielding nutrients, vitamins, minerals, and water do not yield energy in the human body.

Energy Measured in kCalories The energy released from carbohydrate, fat, and protein can be measured in calories—tiny units of energy so small that a single apple provides tens of thousands of them. To ease calculations, energy is expressed in 1000-calorie metric units known as kilocalories (shortened to kcalories, but commonly called "calories"). When you read in popular books or magazines

TABLE 1-1 Elements in the Six Classes of Nutrients

Notice that organic nutrients contain carbon.

	Carbon	Hydrogen	Oxygen	Nitrogen	Minerals
Inorganic nutrients					
Minerals					✓
Water		✓	✓		
Organic nutrients					
Carbohydrate	✓	✓	✓		
Lipid (fat)	✓	✓	✓		
Proteina	✓	✓	✓	✓	
Vitaminsb	✓	✓	✓		

a
Some proteins also contain the mineral sulfur.
b
Some vitamins contain nitrogen; some contain minerals.

♦ Carbohydrate, fat, and protein are sometimes called macronutrients because the body requires them in relatively large amounts (many grams daily). In contrast, vitamins and minerals are micronutrients, required only in small amounts (milligrams or micrograms daily).

inorganic: not containing carbon or pertaining to living things.
• in = not

organic: in chemistry, a substance or molecule containing carbon-carbon bonds or carbon-hydrogen bonds. This definition excludes coal, diamonds, and a few carbon-containing compounds that contain only a

single carbon and no hydrogen, such as carbon dioxide (CO_2), calcium carbonate ($CaCO_3$), magnesium carbonate ($MgCO_3$), and sodium cyanide (NaCN).

essential nutrients: nutrients a person must obtain from food because the body cannot make them for itself in sufficient quantity to meet physiological needs; also called indispensable nutrients. About 40 nutrients are currently known to be essential for human beings.
energy-yielding nutrients: the nutrients that break down to yield energy the body can use:
• Carbohydrate
• Fat

• Protein

calories: units by which energy is measured. Food energy is measured in kilocalories (1000 calories equal

1 kilocalorie), abbreviated kcalories or kcal. One kcalorie is the amount of heat necessary to raise the temperature of 1 kilogram (kg) of water 1°C. The

scientific use of the term kcalorie is the same as the

popular use of the term calorie.

FIGURA 2: Trecho no formato TXT (WHITNEY; ROLFES, 2011; p. 29)
com Página Típica com duas Colunas

Na figura 1, os parágrafos correspondentes às seções de interesse no capítulo, são precedidos pelo nome de suas seções, respectivamente nas cores vermelho, com fonte maior, e azul, com fonte menor. Esta diferenciação visual desaparece no formato TXT, do mesmo trecho da página, ilustrado na figura 2, na qual não existe a diferenciação de cores, nem do tamanho das letras, para separar os nomes das seções das sentenças dos seus respectivos parágrafos. Este problema é rotulado como "Problema 1: separação dos títulos das seções do primeiro parágrafo da seção". Adicionalmente, pode-se observar

que as sentenças dos parágrafos de interesse não estão justicadas no formato TXT, da mesma forma que aparecem no formato PDF, de forma que cada linha tem um tamanho potencialmente diferente.

Entremeando textos de seções e definições aparecem componentes que misturam figuras, tabelas e textos: *Simple Figure* (figuras 3 e 5), Figure (figura 7), *Table* (figuras 1, 9 e 10), e *HowTo_TryIt* (figura 9). A representação destes componentes pode ocupar ou invadir a coluna ocupada pelas seções do capítulo, em uma determinada página, dificultando sensivelmente a separação dos vários grupos de informações.

Na figura 3 é ilustrado um trecho da página, no formato PDF, na qual um componente *Simple Figure* se mantém confinado na sua coluna, sem invadir a coluna lateral. Diferentemente do componente *Figure*, o componente *Simple Figure* é focado na visualização de uma figura, que aparece sempre na coluna das referências, não é demarcado com nenhuma palavra chave, e está associado a dois textos: um texto de *copyright*, na lateral da figura; e um texto localizado logo abaixo da figura.

© Cobas Premium RF/Alamy

An enjoyable way to learn about other cultures is to taste their ethnic foods.

Consumer emphasis on convenience limits food choices to the selections offered on menus and products designed for quick preparation. Whether decisions based on convenience meet a person's nutrition needs depends on the choices made. Eating a banana or a candy bar may be equally convenient, but the fruit provides more vitamins and minerals and less sugar and fat.

Rising food costs have shifted some consumers' priorities and changed their shopping habits. They are less likely to buy higher priced convenience foods and more likely to buy less-expensive store brand items and prepare home-cooked meals. In fact, more than 80 percent of U.S. consumers are eating home-cooked meals at least three times a week.[4]

Positive and Negative Associations People tend to like particular foods associated with happy occasions—such as hot dogs at ball games or cake and ice cream at birthday parties. By the same token, people can develop aversions and dislike foods that they ate when they felt sick or that they were forced to eat as a child. By using foods as rewards or punishments, parents may inadvertently teach their children to like and dislike certain foods.

FIGURA 3: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 26)
com Figura ao Lado a Área de Texto

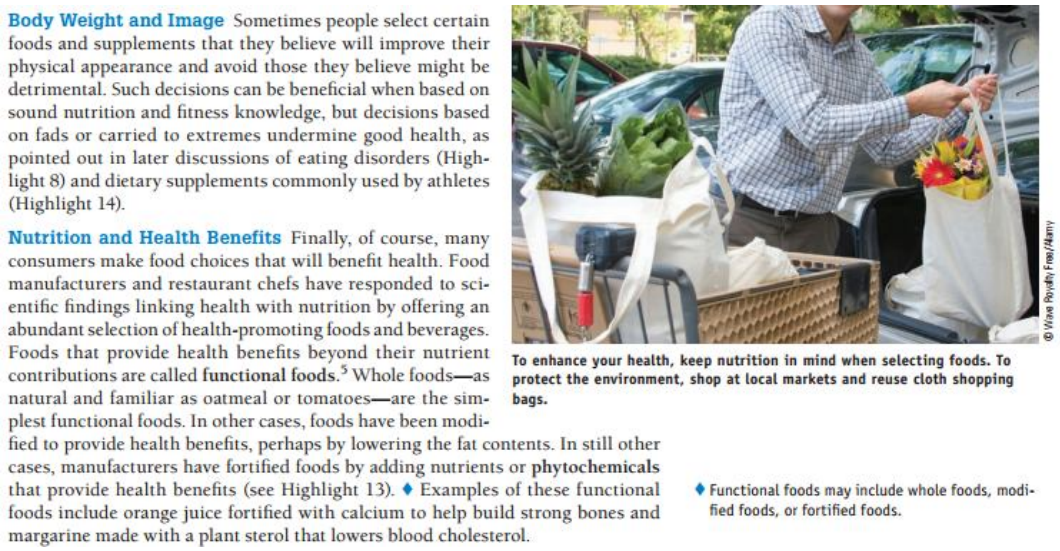
```
                                    Consumer emphasis on convenience limits food choices to the selections of-
                                fered on menus and products designed for quick preparation. Whether decisions
                                based on convenience meet a person's nutrition needs depends on the choices
                                made. Eating a banana or a candy bar may be equally convenient, but the fruit
                                provides more vitamins and minerals and less sugar and fat.
                                    Rising food costs have shifted some consumers' priorities and changed their
                                shopping habits. They are less likely to buy higher priced convenience foods and
                                more likely to buy less-expensive store brand items and prepare home-cooked
y
m
a
l
A
/
F
R
m
u
i
m
e
r
P
s
l
b
r
o
C
                                meals. In fact, more than 80 percent of U.S. consumers are eating home-cooked

                                meals at least three times a week.4




                                Positive and Negative Associations  People tend to like particular foods as-

                                sociated with happy occasions—such as hot dogs at ball games or cake and ice
                                cream at birthday parties. By the same token, people can develop aversions and
                                dislike foods that they ate when they felt sick or that they were forced to eat as a
                                child. By using foods as rewards or punishments, parents may inadvertently teach
                                their children to like and dislike certain foods.
An enjoyable way to learn about other cultures is
to taste their ethnic foods.
```

FIGURA 4: Trecho no formato TXT (WHITNEY; ROLFES, 2011; p. 26)
com Figura ao Lado da Área de Texto

Na figura 4, é ilustrado o mesmo trecho de página no formato TXT. É possível observar, na figura 4, que tanto o texto de *copyright*, ao lado direito da figura do componente *Simple Figure*, bem como seu texto inferior, se mantém confinados na coluna esquerda, sem invadir a coluna direita, que

contém o texto das sentenças de interesse para a extração. Nota-se que as linhas do texto da coluna direita são espaçadas de forma desigual, quando estão posicionadas ao lado da figura do componente *Simple Figure*.

Mas situação ilustrada nas figuras 3 e 4, na qual o componente *Simple Figure* não invade a coluna lateral das sentenças de interesse, não se verifica em muitos casos. Na figura 5, é ilustrado um trecho de página no formato PDF, no qual o componente *Simple Figure* invade a coluna das sentenças.



Body Weight and Image Sometimes people select certain foods and supplements that they believe will improve their physical appearance and avoid those they believe might be detrimental. Such decisions can be beneficial when based on sound nutrition and fitness knowledge, but decisions based on fads or carried to extremes undermine good health, as pointed out in later discussions of eating disorders (Highlight 8) and dietary supplements commonly used by athletes (Highlight 14).

Nutrition and Health Benefits Finally, of course, many consumers make food choices that will benefit health. Food manufacturers and restaurant chefs have responded to scientific findings linking health with nutrition by offering an abundant selection of health-promoting foods and beverages. Foods that provide health benefits beyond their nutrient contributions are called **functional foods**.[5] Whole foods—as natural and familiar as oatmeal or tomatoes—are the simplest functional foods. In other cases, foods have been modified to provide health benefits, perhaps by lowering the fat contents. In still other cases, manufacturers have fortified foods by adding nutrients or **phytochemicals** that provide health benefits (see Highlight 13). ♦ Examples of these functional foods include orange juice fortified with calcium to help build strong bones and margarine made with a plant sterol that lowers blood cholesterol.

© Vitaw Royalty Free/Alamy

To enhance your health, keep nutrition in mind when selecting foods. To protect the environment, shop at local markets and reuse cloth shopping bags.

♦ Functional foods may include whole foods, modified foods, or fortified foods.

FIGURA 5: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 27)
com Figura Invadindo a Área de Texto

FIGURA 6: Trecho no formato TXT (WHITNEY; ROLFES, 2011; p. 27)
com Figura Invadindo a Área de Texto

Na figura 6, é ilustrado o mesmo trecho de página no formato TXT, caracterizando a mistura de texto entre as duas colunas. Este problema é rotulado como “Problema 2: invasão da coluna das sentenças de interesse devido ao componente *Simple Figure* da coluna lateral”.

Na figura 7, é ilustrado um trecho de página no formato PDF, no qual o componente *Figure*, com início delimitado pela palavra chave "*FIGURE*" e com término não demarcado por nenhuma palavra chave, está ocupando a coluna direita, compartilhada com as sentenças de interesse. Na figura 8, é ilustrado o mesmo trecho de página no formato TXT. Nesta ilustração, o componente *Figure* é composto de um texto superior, de uma figura ao centro, tendo um texto de *copyright* à sua direita, e de um texto inferior. Se o texto inferior e o superior trocassem de posição, o único padrão para identificar a fronteira entre o texto inferior do componente Figure, e o parágrafo com sentenças de interesse, seria uma linha em branco. Mas este padrão não é bom discriminante de fronteira, dado que ocorrem situações com linhas em branco entre trechos de texto de um dado componente. Este problema é rotulado como "Problema 3: determinação da fronteira inferior de componentes iniciados por palavras chaves e finalizados sem demarcação".

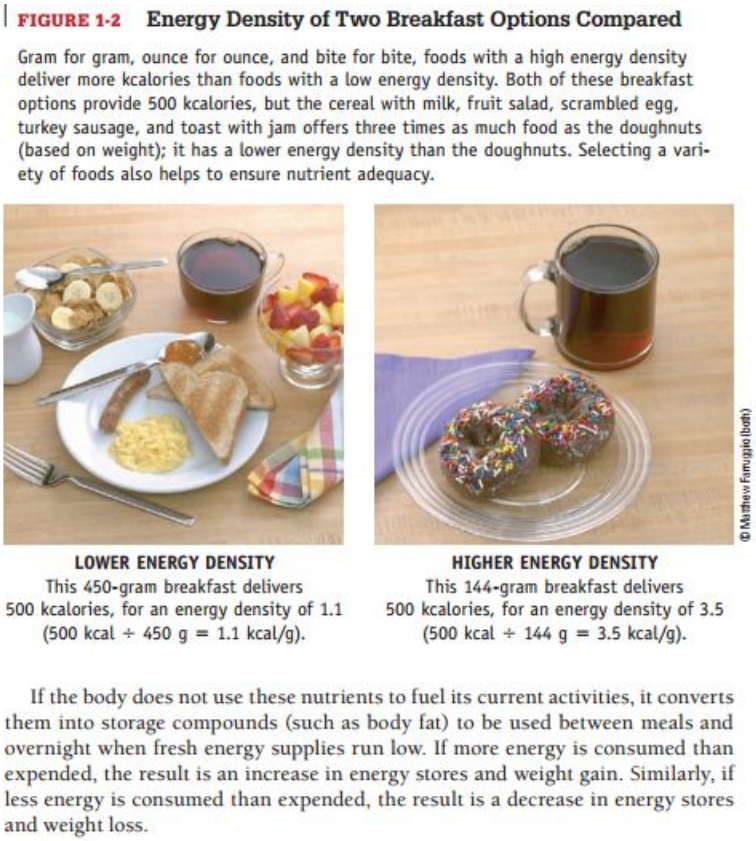


FIGURE 1-2 **Energy Density of Two Breakfast Options Compared**

Gram for gram, ounce for ounce, and bite for bite, foods with a high energy density deliver more kcalories than foods with a low energy density. Both of these breakfast options provide 500 kcalories, but the cereal with milk, fruit salad, scrambled egg, turkey sausage, and toast with jam offers three times as much food as the doughnuts (based on weight); it has a lower energy density than the doughnuts. Selecting a variety of foods also helps to ensure nutrient adequacy.

LOWER ENERGY DENSITY
This 450-gram breakfast delivers 500 kcalories, for an energy density of 1.1 (500 kcal ÷ 450 g = 1.1 kcal/g).

HIGHER ENERGY DENSITY
This 144-gram breakfast delivers 500 kcalories, for an energy density of 3.5 (500 kcal ÷ 144 g = 3.5 kcal/g).

If the body does not use these nutrients to fuel its current activities, it converts them into storage compounds (such as body fat) to be used between meals and overnight when fresh energy supplies run low. If more energy is consumed than expended, the result is an increase in energy stores and weight gain. Similarly, if less energy is consumed than expended, the result is a decrease in energy stores and weight loss.

FIGURA 7: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 32)
com Componente na Mesma Coluna das Sentenças de Interesse

Na figura 9, é ilustrado um trecho de página no formato PDF, no qual o componente *HowTo_TryIt*, com início delimitado pelas palavras chaves "*HOW TO*" e final delimitado pelas palavras chaves "*TRY IT*", se posiciona na coluna esquerda, compartilhada com as sentenças de interesse. Na coluna direita, existe um componente *Table*, com início delimitado pela palavra chave "*TABLE*" e final não delimitado por nenhuma palavra chave, que compartilha esta coluna com

definições. Na figura 10, é ilustrado um trecho de página no formato PDF, no qual o componente *Table* ocupa as duas colunas. De maneira geral, com exceção do componente *Simple Figure*, componentes podem ocupar: qualquer uma das duas colunas, colunas lado a lado ou as duas colunas. Este problema, que aumenta a complexidade do tratamento de componentes, caracterizada inicialmente pelo "Problema 3", é rotulado como "Problema 4: multiplicidade de situações de ocupação de colunas por componentes".

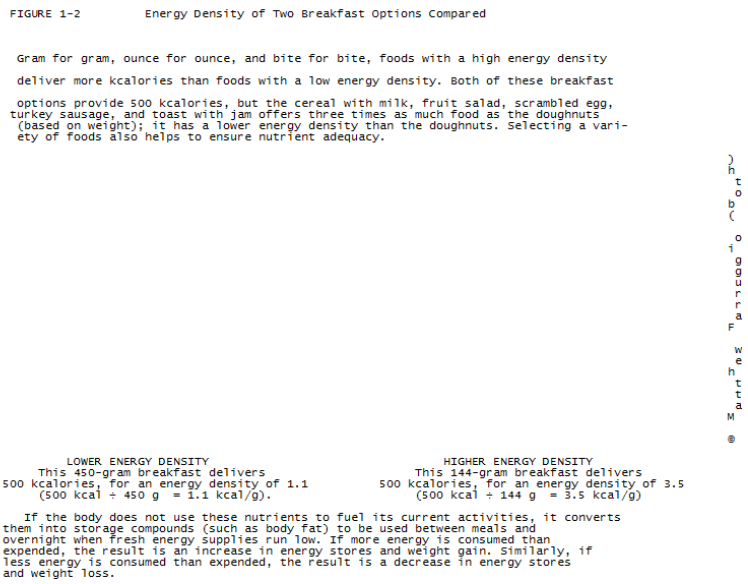


FIGURE 1-2 Energy Density of Two Breakfast Options Compared

Gram for gram, ounce for ounce, and bite for bite, foods with a high energy density deliver more kcalories than foods with a low energy density. Both of these breakfast options provide 500 kcalories, but the cereal with milk, fruit salad, scrambled egg, turkey sausage, and toast with jam offers three times as much food as the doughnuts (based on weight); it has a lower energy density than the doughnuts. Selecting a variety of foods also helps to ensure nutrient adequacy.

LOWER ENERGY DENSITY
This 450-gram breakfast delivers 500 kcalories, for an energy density of 1.1 (500 kcal ÷ 450 g = 1.1 kcal/g).

HIGHER ENERGY DENSITY
This 144-gram breakfast delivers 500 kcalories, for an energy density of 3.5 (500 kcal ÷ 144 g = 3.5 kcal/g)

If the body does not use these nutrients to fuel its current activities, it converts them into storage compounds (such as body fat) to be used between meals and overnight when fresh energy supplies run low. If more energy is consumed than expended, the result is an increase in energy stores and weight gain. Similarly, if less energy is consumed than expended, the result is a decrease in energy stores and weight loss.

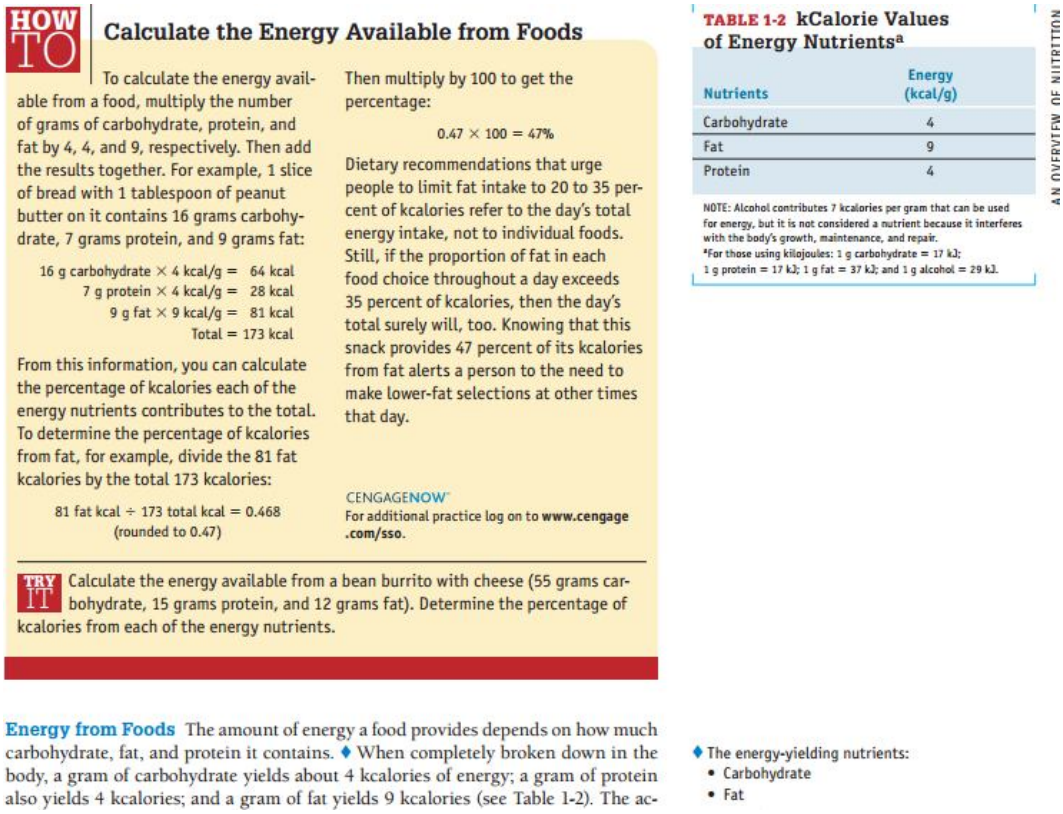
FIGURA 8: Trecho no formato TXT (WHITNEY; ROLFES, 2011; p. 32)
com Componente na Mesma Coluna das Sentenças de Interesse

Existem situações nas quais textos, de uma dada linha, que no formato PDF pertencem a colunas distintas, ao serem convertidos para o formato TXT, se misturam de forma a não ser possível separá-los em suas colunas originais. Na figura 11, esta situação é ilustrada no formato PDF, no qual a divisão das colunas é bastante clara. Na figura 12, é ilustrado o mesmo trecho de página no formato TXT, no qual os textos se misturam em algumas linhas. Este problema é rotulado como "Problema 5: mistura de texto oriundos de colunas distintas".

Finalmente, *bullets* são utilizados, no texto das sentenças de interesse, em várias situações: para indentar conjuntos de palavras ou conjunto de sentenças. Na figura 13, é ilustrada, no formato PDF, a utilização de *bullets* para indentar conjuntos de palavras distintos em uma mesma linha.

De forma equivalente, a indentação de sentenças de interesse, pode ser utilizada em parágrafos enumerados. Na figura 14, é ilustrada, no formato PDF, a utilização de indentação em parágrafos enumerados. Pode-se concluir que a utilização de indentação através de *bullets* ou enumerados é representada de várias formas no texto: conjuntos de palavras indentados por *bullets*, em linhas separadas ou na mesma linha; e sentenças e paráfrafos indentados por *bullets* ou enumerados. Estas

situações precisam ser tratadas para evitar a perda da informação de indentação do texto original. Este problema é rotulado como “Problema 6: indentação através de *bullets* ou enumerados”.



HOW TO

Calculate the Energy Available from Foods

To calculate the energy available from a food, multiply the number of grams of carbohydrate, protein, and fat by 4, 4, and 9, respectively. Then add the results together. For example, 1 slice of bread with 1 tablespoon of peanut butter on it contains 16 grams carbohydrate, 7 grams protein, and 9 grams fat:

16 g carbohydrate × 4 kcal/g = 64 kcal
7 g protein × 4 kcal/g = 28 kcal
9 g fat × 9 kcal/g = 81 kcal
Total = 173 kcal

From this information, you can calculate the percentage of kcalories each of the energy nutrients contributes to the total. To determine the percentage of kcalories from fat, for example, divide the 81 fat kcalories by the total 173 kcalories:

81 fat kcal ÷ 173 total kcal = 0.468 (rounded to 0.47)

Then multiply by 100 to get the percentage:

0.47 × 100 = 47%

Dietary recommendations that urge people to limit fat intake to 20 to 35 percent of kcalories refer to the day's total energy intake, not to individual foods. Still, if the proportion of fat in each food choice throughout a day exceeds 35 percent of kcalories, then the day's total surely will, too. Knowing that this snack provides 47 percent of its kcalories from fat alerts a person to the need to make lower-fat selections at other times that day.

CENGAGENOW
For additional practice log on to **www.cengage.com/sso**.

TRY IT Calculate the energy available from a bean burrito with cheese (55 grams carbohydrate, 15 grams protein, and 12 grams fat). Determine the percentage of kcalories from each of the energy nutrients.

TABLE 1-2 kCalorie Values of Energy Nutrients[a]

Nutrients	Energy (kcal/g)
Carbohydrate	4
Fat	9
Protein	4

NOTE: Alcohol contributes 7 kcalories per gram that can be used for energy, but it is not considered a nutrient because it interferes with the body's growth, maintenance, and repair.
[a]For those using kilojoules: 1 g carbohydrate = 17 kJ; 1 g protein = 17 kJ; 1 g fat = 37 kJ; and 1 g alcohol = 29 kJ.

AN OVERVIEW OF NUTRITION

Energy from Foods The amount of energy a food provides depends on how much carbohydrate, fat, and protein it contains. ♦ When completely broken down in the body, a gram of carbohydrate yields about 4 kcalories of energy; a gram of protein also yields 4 kcalories; and a gram of fat yields 9 kcalories (see Table 1-2). The ac-

♦ The energy-yielding nutrients:
- Carbohydrate
- Fat

FIGURA 9: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 31) com componente *HowTo_TryIt* na coluna esquerda e o componente *Table* na coluna direita

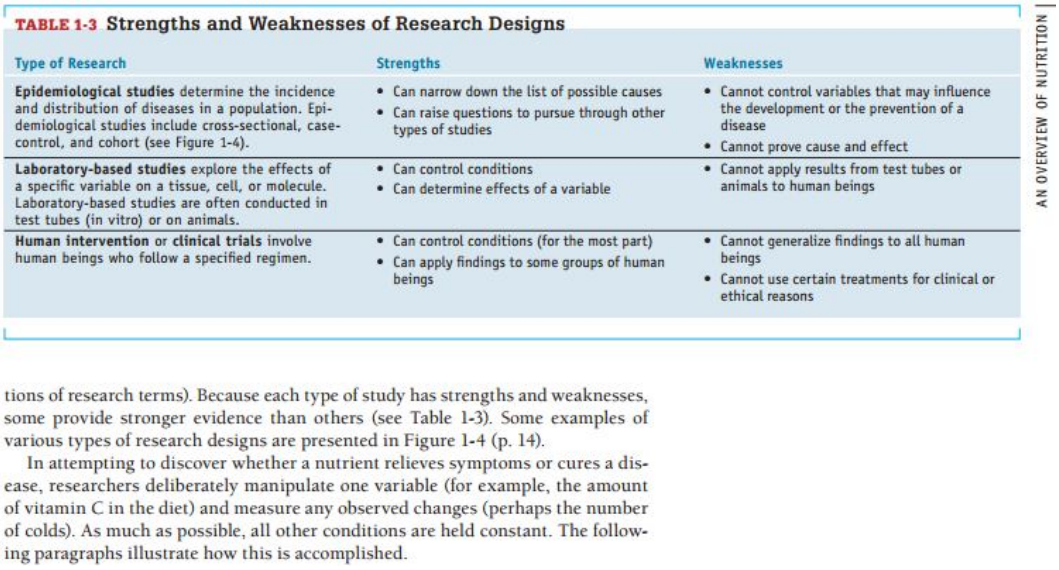


TABLE 1-3 Strengths and Weaknesses of Research Designs

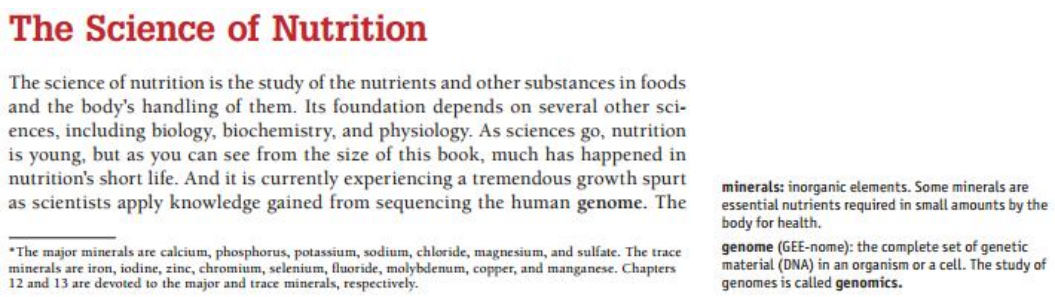
Type of Research	Strengths	Weaknesses
Epidemiological studies determine the incidence and distribution of diseases in a population. Epidemiological studies include cross-sectional, case-control, and cohort (see Figure 1-4).	• Can narrow down the list of possible causes • Can raise questions to pursue through other types of studies	• Cannot control variables that may influence the development or the prevention of a disease • Cannot prove cause and effect
Laboratory-based studies explore the effects of a specific variable on a tissue, cell, or molecule. Laboratory-based studies are often conducted in test tubes (in vitro) or on animals.	• Can control conditions • Can determine effects of a variable	• Cannot apply results from test tubes or animals to human beings
Human intervention or **clinical trials** involve human beings who follow a specified regimen.	• Can control conditions (for the most part) • Can apply findings to some groups of human beings	• Cannot generalize findings to all human beings • Cannot use certain treatments for clinical or ethical reasons

AN OVERVIEW OF NUTRITION

tions of research terms). Because each type of study has strengths and weaknesses, some provide stronger evidence than others (see Table 1-3). Some examples of various types of research designs are presented in Figure 1-4 (p. 14).

In attempting to discover whether a nutrient relieves symptoms or cures a disease, researchers deliberately manipulate one variable (for example, the amount of vitamin C in the diet) and measure any observed changes (perhaps the number of colds). As much as possible, all other conditions are held constant. The following paragraphs illustrate how this is accomplished.

FIGURA 10: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 35) com componente *Table* Ocupando as duas Colunas



The Science of Nutrition

The science of nutrition is the study of the nutrients and other substances in foods and the body's handling of them. Its foundation depends on several other sciences, including biology, biochemistry, and physiology. As sciences go, nutrition is young, but as you can see from the size of this book, much has happened in nutrition's short life. And it is currently experiencing a tremendous growth spurt as scientists apply knowledge gained from sequencing the human **genome**. The

*The major minerals are calcium, phosphorus, potassium, sodium, chloride, magnesium, and sulfate. The trace minerals are iron, iodine, zinc, chromium, selenium, fluoride, molybdenum, copper, and manganese. Chapters 12 and 13 are devoted to the major and trace minerals, respectively.

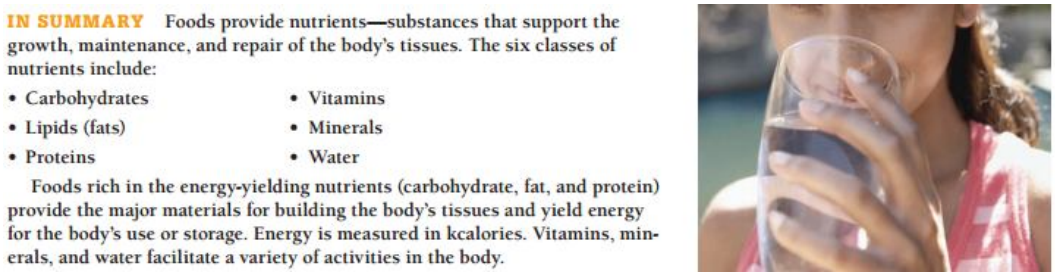
minerals: inorganic elements. Some minerals are essential nutrients required in small amounts by the body for health.

genome (GEE-nome): the complete set of genetic material (DNA) in an organism or a cell. The study of genomes is called **genomics.**

FIGURA 11: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 33)
com Nota de Rodapé ao Lado de Definições

```
The Science of Nutrition

The science of nutrition is the study of the nutrients and other substances in foods
and the body's handling of them. Its foundation depends on several other sci-
ences, including biology, biochemistry, and physiology. As sciences go, nutrition
is young, but as you can see from the size of this book, much has happened in
nutrition's short life. And it is currently experiencing a tremendous growth spurt
as scientists apply knowledge gained from sequencing the human genome. The

                                                    minerals: inorganic elements. Some minerals are
                                                    essential nutrients required in small amounts by the
                                                    body for health.

*The major minerals are calcium, phosphorus, potassium, sodium, chloride, magnesium, and sulfate. The trace genome (GEE-nome): the complete set of genetic
minerals are iron, iodine, zinc, chromium, selenium, fluoride, molybdenum, copper, and manganese. Chapters material (DNA) in an organism or a cell. The study of
12 and 13 are devoted to the major and trace minerals, respectively.                genomes is called genomics.
```

FIGURA 12: Trecho no formato TXT (WHITNEY; ROLFES, 2011; p. 33)
com Mistura de Texto entre Nota de Rodapé e Definições



IN SUMMARY Foods provide nutrients—substances that support the growth, maintenance, and repair of the body's tissues. The six classes of nutrients include:

- Carbohydrates
- Lipids (fats)
- Proteins
- Vitamins
- Minerals
- Water

Foods rich in the energy-yielding nutrients (carbohydrate, fat, and protein) provide the major materials for building the body's tissues and yield energy for the body's use or storage. Energy is measured in kcalories. Vitamins, minerals, and water facilitate a variety of activities in the body.

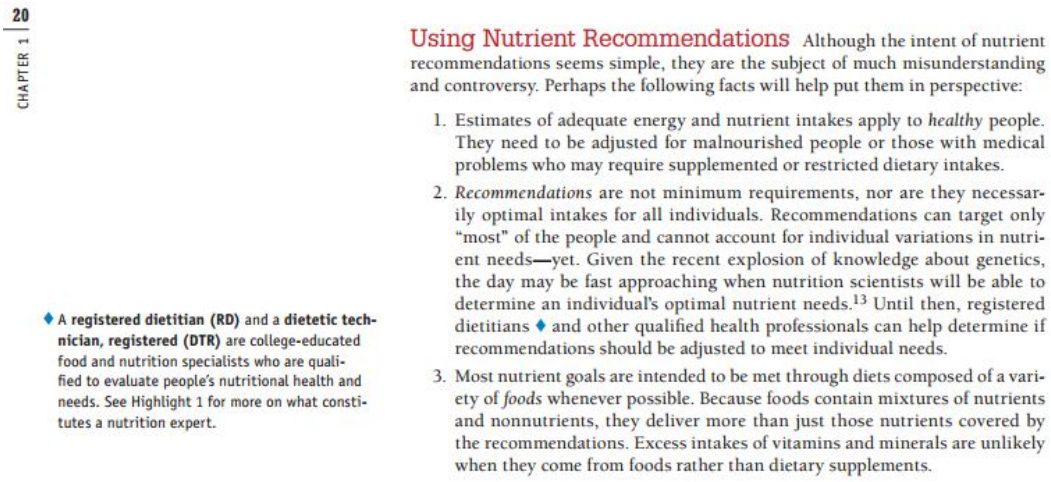
FIGURA 13: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 33)
com Mais de um *Bullet* na Mesma Linha

Nesta seção, foram ilustrados e identificados os principais problemas tratados pelo protótipo, proposto neste trabalho. Existem vários outros problemas menores, que precisaram ser tratados na implementação do protótipo como, por exemplo, referências numéricas que aparecem ao final de algumas linhas no formato PDF, mas que na conversão para o formato TXT, passam a ocupar isoladamente a linha inferior de sua linha de origem. Foi necessário selecionar os problemas mais relevantes para abordar neste artigo, pois não existe espaço suficiente para abordar neste trabalho todos os problemas, e nem as variações de alguns tipos de problemas, que precisaram ser tratadas na implementação do protótipo.

1.2 SOLUÇÃO PARA OS PROBLEMAS IDENTIFICADOS

A segunda etapa do trabalho é a proposta da solução para resolver os problemas, identificados na seção anterior, que dificultam o isolamento do texto das sentenças de interesse para extração: (1) separação dos títulos das seções do primeiro parágrafo da seção; (2) invasão da coluna das sentenças de

interesse devido ao componente *Simple Figure* da coluna lateral; (3) determinação da fronteira inferior de componentes iniciados por palavras chaves e finalizados sem demarcação; (4) multiplicidade de situações de ocupação de colunas por componentes; (5) mistura de textos oriundos de colunas distintas; e (6) indentação através de *bullets* ou enumerados.



20 | CHAPTER 1

♦ A **registered dietitian (RD)** and a **dietetic technician, registered (DTR)** are college-educated food and nutrition specialists who are qualified to evaluate people's nutritional health and needs. See Highlight 1 for more on what constitutes a nutrition expert.

Using Nutrient Recommendations Although the intent of nutrient recommendations seems simple, they are the subject of much misunderstanding and controversy. Perhaps the following facts will help put them in perspective:

1. Estimates of adequate energy and nutrient intakes apply to *healthy* people. They need to be adjusted for malnourished people or those with medical problems who may require supplemented or restricted dietary intakes.
2. *Recommendations* are not minimum requirements, nor are they necessarily optimal intakes for all individuals. Recommendations can target only "most" of the people and cannot account for individual variations in nutrient needs—yet. Given the recent explosion of knowledge about genetics, the day may be fast approaching when nutrition scientists will be able to determine an individual's optimal nutrient needs.[13] Until then, registered dietitians ♦ and other qualified health professionals can help determine if recommendations should be adjusted to meet individual needs.
3. Most nutrient goals are intended to be met through diets composed of a variety of *foods* whenever possible. Because foods contain mixtures of nutrients and nonnutrients, they deliver more than just those nutrients covered by the recommendations. Excess intakes of vitamins and minerals are unlikely when they come from foods rather than dietary supplements.

FIGURA 14: Trecho no formato PDF (WHITNEY; ROLFES, 2011; p. 42) com Parágrafos Enumerados

A solução para os problemas, identificados na seção anterior, pode ser decomposta em cinco passos: (a) identificação dos intervalos de linhas ocupadas por componentes em cada página; (b) separação dos textos associados às colunas direita e esquerda em cada página; (c) concatenação dos textos associados à coluna de interesse em cada página; (d) identificação das palavras, conjuntos de palavras, sentenças e parágrafos indentadas por bullets e enumerados; e (e) identificação dos títulos das seções que se misturam com o texto do primeiro parágrafo de uma dada seção.

O primeiro passo da solução proposta é a identificação dos intervalos de linhas ocupadas por componentes em cada página. A execução completamente automatizada desta atividade não é simples, em função dos seguintes problemas: multiplicidade de situações de ocupação de colunas por componentes; e dificuldade de determinar a fronteira inferior de componentes iniciados por palavras chaves e finalizados sem demarcação. A solução encontrada foi introduzir, manualmente, as seguintes marcas no arquivo com formato TXT: (a) a palavra chave "*ALL*" após as palavras chaves "*FIGURE*" e "*TABLE*" dos respectivos componentes que ocupam as duas colunas da página; e (b) as palavras chaves "*ENDTABLE*" e "*ENDFIGURE*" para demarcar a linha em branco que sucede o final dos componentes *Table* e *Figure*, cujo final não coincida com o final da página.

O segundo passo da solução proposta é a separação dos textos associados às colunas direita e esquerda de cada página. Conforme comentado, na seção anterior, as colunas ocupadas pelas seções do capítulo e pelas definições se alternam a cada página. Aparentemente, seria possível determinar um valor fixo de coluna para as páginas impares, e outro para as páginas pares, que demarquem a fronteiras

entre as colunas direita e esquerda, nas páginas pares e nas página impares. Essa solução simplista não é geral por dois motivos: (a) o intervalo de colunas que separa as duas colunas varia ao longo das páginas do capítulo; e (b) com base no problema "mistura de textos oriundos de colunas distintas", não é possível separar automaticamente as duas colunas em algumas situações de mistura de texto.

A solução encontrada foi a determinação automática, para cada página específica, do intervalo de colunas que serve de fronteira para as colunas esquerda e direita da respectiva página. Palavras são ligadas por um espaço nas sentenças. Espaçamentos entre as colunas direita e esquerda precisam ser caracterizados por pelo menos dois espaços. As fronteiras das páginas variam, mas ocorrem dentro de um determinado intervalo de valores para as páginas ímpares, e em outro intervalo de valores para as páginas pares. Nestes intervalos de valores é determinado o intervalo de colunas que não se sobrepõe com nenhum texto na página. Embora esta solução possa parecer simples, a sua implementação não é simples em função da multiplicidade de situações que ocorrem no formato TXT das páginas do corpus. Na figura 2 é possível perceber que os dois intervalos de colunas que rodeam a palavra "*Minerals*" não se sobrepõe com os textos da página, no entanto, somente o segundo intervalo divide corretamente as duas colunas da página. Na figura 4, pode-se observar que o intervalo de colunas, que serve de fronteira para as colunas direita e esquerda, está localizado em um curto espaço entre o copyright da figura à esquerda e as sentenças à direita.

Adicionalmente, esta solução inicial precisa ser aperfeiçoada para tratar as situações caracterizadas por dois problemas. O primeiro problema, "invasão da coluna das sentenças de interesse devido ao componente *Simple Figure* da coluna lateral", caracteriza a necessidade de definir um intervalo de colunas para o primeiro intervalo de linhas do trecho ilustrado, e um outro intervalo de colunas para o segundo intervalo de linhas do trecho ilustrado. Na figura 6, é possível observar que a linha ocupada pelo texto "*fied to provide health benefits, perhaps by lowering the fat contents. In still other*" bloqueia o intervalo de colunas de fronteita limitado pelo espaço existente entre as palavras "*nutrient*" e "*To*", resultando na determinação de um intervalo de colunas de fronteira adicional, a partir da referida linha, que passa a ser limitado pelo espaço existente entre a palavra "*other*" e o símbolo "♦".

O segundo problema, "mistura de textos oriundos de colunas distintas", ilustrado na figura 12, caracteriza a necessidade de idenficar uma separação entre as palavras contíguas de sentenças distintas em uma dada linha. No caso da separação entre as palavras "*Chapter*" e "*material*" é possível identificar que a palavra "*Chapter*" ficará na coluna esquerda porque se sobrepõe com a coluna mais à direita do intervalo de colunas que serve como fronteira. No entanto, esta solução não funciona na separação entre as palavras "*trace*" e "*genome*". A solução para esta situação muito específica, que ocorre com rara frequência em cada capítulo do corpus, foi introduzir manualmente um marcador para caracterizar a separação entre as palavras: "*trace*|*genome*".

O terceiro passo da solução proposta é a concatenação dos textos associados à coluna de interesse em cada página. Para cada capítulo são concatenados os trechos de texto, desconsiderando as linhas ocupadas por componentes, de cada página que contém as seções do capítulo. Conforme já comentado, a cada página as seções do capítulo ocupam a coluna esquerda ou a coluna direita da página, alternadamente. O resultado final é um único conjunto de linhas para todo o capítulo. As referências ao final do capítulo não são consideradas.

O quarto passo da solução proposta é o isolamento das palavras, conjuntos de palavras, sentenças e parágrafos indentadas por *bullets* e enumerados. Para todos estes casos é necessário: identificar a existência do *bullet* ou enumerado; e identificar qual a sua próxima ocorrência ou o término de sua ocorrência. Se os elementos que utilizam *bullets*, são palavras ou conjuntos de palavras, com um ou mais *bullets* na mesma linha, estes elementos são encadeados, na sentença, separados por ponto e vírgula. Se os elementos que utilizam *bullets* ou enumerados são sentenças ou parágrafos, estes elementos serão concatenados, sem o *bullet* ou enumerado, para posterior extração das sentenças.

O quinto passo da solução proposta é a remoção dos títulos das seções que se misturam com o texto do primeiro parágrafo de uma dada seção. Na figura 2, é possível o observar que os títulos dos dois parágrafos de interesse se distanciam da primeira sentença do parágrafo por dois espaços, e a partir deste ponto, as palavras da primeira sentença do parágrafo passam a ser separadas por apenas um espaço. Este espaçamento se observa tanto entre o título "*The Energy-Yielding Nutrients: Carbohydrate, Fat, and Protein*" e sentença inicial do parágrafo "*In the body, ...*", como entre o título "*Energy Measured in kCalories*" e a sentença "*The energy released ...*". Desta forma, este espaço serve de fronteira para a separação entre título da seção e a primeira sentença do primeiro parágrafo da seção.

Após a execução do quinto passo da solução proposta, o conjunto de linhas de texto resultante contém somente as sentenças das seções de cada capítulo, de forma semelhante aos documentos que são utilizados como corpus para que detectores de sentença convencionais (APACHE, 2010) possam realizar a extração das sentenças. Desta forma, o resultado da execução do quinto passo, é utilizado para a separação do conjunto de linhas resultante em sentenças, considerando as pontuações de final de sentença como separadores.

1.3 PROTOTIPAÇÃO DA SOLUÇÃO PROPOSTA

A terceira etapa do trabalho é a prototipagem da solução proposta, na seção anterior, para extração dos trechos de texto e de suas sentenças.

Embora o foco principal deste artigo seja a extração das sentenças das seções dos capítulos, a implementação foi estruturada para evoluir para a extração das demais partes do corpus, tais como: as sentenças dos *highlights* (textos complementares de cada capítulo), as definições que ocorrem nas

colunas laterais às seções do capítulo em cada página, a hierarquia de seções de cada capítulo, os glossários de termos que aparecem nos capítulos, as referências ao final de cada capítulo, e inclusive as sentenças que aparecem nos componentes *Figure*, *Table* e *HowTo_TryIt*. Adicionalmente, serão considerados os diferentes tipos referências que aparecem nas sentenças das seções dos capítulos: os números associados a referências no final do capítulo; as referências a textos explicativos nas colunas laterais através do símbolo "♦"; e as referências a textos de rodapé a partir do símbolo "*".

Os problemas identificados e a solução proposta, nas seções anteriores, é uma visão de alto nível das principais questões relavantes para a implementação do protótipo relatado neste trabalho, cuja implementação envolveu outros aspectos e variantes das questões analisadas.

Até o presente momento, foram extraídas corretamente as sentenças das seções dos quatro primeiro capítulos do corpus, que contém ao todo vinte capítulos. As quantidades de sentenças extraídas, para cada um dos quatro capítulos, são respectivamente: 510, 412, 359 e 555.

A cada novo capítulo tratado, aparecem variantes dos problemas identificados anteriormente, cujo tratamento resulta na evolução da implementação corrente. Desta forma, este trabalho relata o estágio atual da implementação do protótipo, cujo objetivo final é a extração correta das sentenças de todos os capítulos do corpus.

CONSIDERAÇÕES FINAIS

A solução proposta, especialmente para a separação entre as colunas das páginas, automatiza um tratamento para uma grande variedade de situações. A caracterização correta das fronteiras entre as colunas esquerda e direita, nas páginas do corpus, não é simples em função das invasões, que os espaços ocupados pelas sentenças das seções dos capítulos sofrem em função da representação dos componentes. Estas invasões de espaço, que não existem originalmente no formato PDF, mas que são geradas na conversão para o formato TXT, caracterizam contornos de fronteira completamente irregulares, entre as colunas esquerda e direita de cada página.

Conforme relatado na seção 1.2, a solução proposta, para a implementação do protótipo apresentado neste trabalho, requer a inserção manual, no arquivo com formato TXT do corpus, de marcadores para viabilizar a detecção: (a) da ocupação de colunas dos componentes *Figure*, *Table* e *HowTo_TryIt*; (b) das linhas de término dos componente *Figure* e *Table*, que não coincidem com o final de página que ocupam; e (c) da separação de palavras de sentenças de colunas distintas que não pode ser realizada de forma automática. Embora a introdução destes marcadores tenha tornado o processo semi-automático, a quantidade de marcadores inseridos foi muito pequena em relação ao texto de entrada. Tomando como referência o primeiro capítulo, foram realizadas: (a) três inserções da palavra chave "*ALL*" para demarcar que um dado componente utiliza as duas colunas da página; (b) dez

inserções da palavra chave "*ENDFIGURE*" ou "*ENDTABLE*" para demarcar o término de um dado componente cujo fim não coincide com o final da página; e (c) quatro inserções do separador "|" para delimitar a separação das sentenças cuja mistura de texto não foi possível tratar automaticamente. Conclui-se que o esforço de introdução manual dos marcadores é pouco significativo em relação ao resultado gerado automaticamente pelo protótipo. A utilização de um processo semi-automático se justifica neste caso, dado que em função das características específicas de distribuição de textos, figuras, tabelas e outros componentes ao longo das páginas do corpus utilizado, este protótipo será utilizado somente para a extração das sentenças deste corpus.

O esforço dispendido para a concepção, implementação e teste do protótipo, desenvolvido neste trabalho, foi muito além do que havia sido estimado inicialmente. Em consequência desta constatação, até o presente momento foi possível extrair corretamente as sentenças dos quatro primeiros capítulos do corpus, que contém ao todo vinte capítulos. O prosseguimento deste trabalho resultará no aperfeiçoamento da solução proposta, para tratar as variantes que surgirem por ocasião do teste de cada novo capítulo do corpus. No entanto, a extração das sentenças de todos os capítulos do corpus é estratégica, como um ponto de partida para pesquisas de construção e extensão automática de ontologias, a partir de uma fonta fidedigna.

A meta, para trabalhos futuros, é utilizar o resultado deste trabalho, direcionado atualmente para a extração de sentenças das seções dos capítulos, para evoluir para a extração e interligação de toda a informação útil do corpus, conforme comentado na seção 1.3, para servir de base para pesquisas em consultas e inferências a partir da informação extraída.

Adicionalmente, a solução proposta neste trabalho será adaptada para a extração de sentenças de artigos técnicos disponíveis na internet. Embora artigos técnicos sejam mais simples do ponto de vista distribuição espacial de texto, tabelas e figuras que utilizam, apresentam um desafio mais complexo, se comparados com o corpus adotado neste trabalho. Neste caso a meta será a implementação de um protótipo capaz de detectar sentenças de artigos técnicos em geral, convertidos para o formato TXT a partir do formato PDF, sem a necessidade da inserção manual de marcadores, a cada novo artigo processado. A restrição, de conceber e implementar um processo completamente automatizado, introduz novos desafios para a evolução deste trabalho.

REFERÊNCIAS

ADOBE, Adobe Reader VI. URL: http://www.adobe.com/products/reader.html. Acesso em 2013.

APACHE, OpenNLP. General Apache OpenNLP Developer Documentation. 2010. URL: http://incubator.apache.org/opennlp/.

ETZIONI, Oren; FADER, A.; CHRISTENSEN, Janara; SODERLAND, Stephen; MAUSAM, Mausam. Open Information Extraction: The Second Generation. In International Joint Conference on Artificial Intelligence, 2011, p. 3-10.

FADER, Anthony; SODERLAND, Stephen; ETZIONI, Oren. Identifying Relations for Open Information Extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, p. 1535–1545.

FOXIT, Foxit Reader 6. URL: http://www.foxitsoftware.com/portuguese/products/reader/. Acesso em 2013.

JOHANSSON, Richard; NUGUES, Pierre. The effect of syntactic representation on semantic role labeling. Proceedings of the 22nd International Conference on Computational Linguistics, 2008, p. 393-400.

KINGSBURY, Paul; PALMER, Martha. From TreeBank to PropBank. 2002, p. 1989-1993.

KINGSBURY, Paul; PALMER, Martha; GILDEA, Daniel. The Proposition Bank: An Annotated Corpus of Semantic Roles. 2004, p. 1-33.

PUNYAKANOK, Vasin; ROTH, Dan; YIH, Wen-tau. The importance of syntactic parsing and inference insemantic role labeling. Computational Linguistics, 34(2), 2008, p. 1-30.

SARAWAGI, Sunita. Information Extraction. Foundations and Trends in Databases: Vol. 1: No 3, 2008, p. 261-377.

SUCHANEK, Fabian M.; KASNECI, Gjergji; WEIKUM, Gerhard. YAGO: A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Websemantics, 2008, p. 203-217.

SUCHANEK, Fabian M.; KASNECI, Gjergji; WEIKUM, Gerhard. SOFIE: A Self-Organizing Framework for Information Extraction. In: Proceedings of the International Conference on World Wide Web, p. 631-640, 2009.

WHITNEY Ellie; ROLFES Sharon Rady. Understanding Nutrition; Twelfth Edition; 2011, 1007 p. Wadsworth Cengage Learning.

WU, Fei; WELD, Daniel S. Automatically refining the Wikipedia infobox ontology. In Proceedings of the International Conference on World Wide Web, p. 635-644, 2008.