Identifying paraphrases between technical and lay corpora

Louise Deléger, Pierre Zweigenbaum

LIMSI – CNRS, Orsay, France

LREC 20th May

Outline

- **Context and Objectives**
- **Building Comparable Corpora**
- **Extracting Paraphrases**
  - Identifying relevant types of paraphrase
  - Extracting nominalization paraphrases
  - Extracting paraphrases of neo-classical compounds
- **Evaluation and Results**
- **Discussion and Conclusion**

Context and Objectives

- **Paraphrase identification** useful to a number of applications
  - IE, QA, text simplification, authoring aids
- Paraphrases between **different discourse types**
  - **technical** vs. **lay** discourse types
  - especially useful for *text simplification* and *authoring aids*
  - in a **medical context** : help produce texts more easily understandable by lay people / patients

treatment of the affection ~ the disease is treated

prostatectomy ~ removal of the prostate

Context and Objectives

→ **Compile a lexicon of paraphrases**

- **specialized phrases** used in **technical texts**
  (for medical specialists)
- **lay phrases** used in **lay texts**
  (for the general public)

Related Work

- **Existing approaches** in paraphrase identification
  - In **parallel corpora** : novel translations [Barzilay & McKeown 2001] bilingual corpora [Max 2009]
  - In **comparable corpora** : newspaper articles [Barzilay & Lee 2003; Shinyama & Sekine 2003], medical corpora [Elhadad & Sutaria 2007]
- Few distinguish **between discourse types**
  - Technical vs. lay discourse types in medical comparable corpora
  - Paraphrase lexicon based on existing English medical terminologies [Elhadad & Sutaria 2007]
  - Identification of relevant types of paraphrases in French medical corpora [Deléger & Zweigenbaum 2009]

→ **Adapt to English an approach originally designed for French** (Deléger & Zweigenbaum 2009)

Building Comparable Corpora

- **Technical vs. lay discourse types**
  - Texts for **medical specialists** vs. texts for **the general public**
- **Two medical topics**
  - Diabetes, Cancer
- **Methods**
  - Website with comparable documents (Cancer corpus)

    Guidelines for physicians vs. lay versions for the general public
  - Guided search through the web (Diabetes corpus)

    Scientific articles (MEDLINE), online medical manual (Merck), websites for the general public (WebMD, NetDoctor), etc.

Technical vs. Lay Documents

Technical

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Glucose Control and Vascular Complications in Veterans with Type 2 Diabetes

William Duckworth, M.D., Carlos Abraira, M.D., Thomas Moritz, M.S., Domenic Reda, Ph.D., Nicholas Emanuele, M.D., Peter D. Reaven, M.D., Franklin J. Zieve, M.D., Ph.D., Jennifer Marks, M.D., Stephen N. Davis, M.D., Rodney Hayward, M.D., Stuart R. Warren, J.D., Pharm.D., Steven Goldman, M.D., Madeline McCarren, Ph.D., M.P.H., Mary Ellen Vitek, William G. Henderson, Ph.D., and Grant D. Huang, M.P.H., Ph.D., for the VADT Investigators*

ABSTRACT

BACKGROUND

The effects of intensive glucose control on cardiovascular events in patients with long-standing type 2 diabetes mellitus remain uncertain.

METHODS

We randomly assigned 1791 military veterans (mean age, 60.4 years) who had a suboptimal response to therapy for type 2 diabetes to receive either intensive or standard glucose control. Other cardiovascular risk factors were treated uniformly. The mean number of years since the diagnosis of diabetes was 11.5, and 40% of the patients had already had a cardiovascular event. The goal in the intensive-therapy group was an absolute reduction of 1.5 percentage points in the glycated hemoglobin level, as compared with the standard-therapy group. The primary outcome was the time from randomization to the first occurrence of a major cardiovascular event, a composite of myocardial infarction, stroke, death from cardiovascular causes, congestive heart failure, surgery for vascular disease, inoperable coronary disease, and amputation for ischemic gangrene.

RESULTS

The median follow-up was 5.6 years. Median glycated hemoglobin levels were 8.4% in the standard-therapy group and 6.9% in the intensive-therapy group. The primary outcome occurred in 264 patients in the standard-therapy group and 235 patients in the intensive-therapy group (hazard ratio in the intensive-therapy group, 0.88; 95% confidence interval [CI], 0.74 to 1.05; P=0.14). There was no significant difference between the two groups in any component of the primary outcome or in the rate of death from any cause (hazard ratio, 1.07; 95% CI, 0.81 to 1.42; P=0.62). No differences between the two groups were observed for microvascular complications. The rates of adverse events, predominantly hypoglycemia, were 17.6% in the standard-therapy group and 24.1% in the intensive-therapy group.

CONCLUSIONS

Intensive glucose control in patients with poorly controlled type 2 diabetes had no significant effect on the rates of major cardiovascular events, death, or microvascular complications, with the exception of progression of albuminuria (P=0.01). (ClinicalTrials.gov number, NCT00032487.)

Lay

Diabetes Guide

Overview & Facts

Understanding diabetes is the first step to managing it. Get information on diabetes causes, risk factors, warning signs, and prevention tips.

FONT SIZE A A A

What Is Diabetes?

Diabetes Overview

Find information on diabetes from the National Institute of Diabetes and Digestive and Kidney Diseases.

Causes

What Are the Causes of Diabetes?

Diabetes occurs when the body cannot regulate blood sugars. Are the causes for type 1 and type 2 diabetes different? Find out.

Are You at Risk?

Are You at Risk for Diabetes?

A family history and age increases the risk for type 2 diabetes. Read what other risk you have; you may be surprised to find that there are some risk factors for type 2 diabetes that you can change.

Are You at Risk for Gestational Diabetes?

Learn about risk factors for diabetes in pregnancy.

Prevention

Diabetes Prevention

Type 1 diabetes can't be prevented, but type 2 diabetes has modifiable risk factors which can help you lower your risk for the disease. Find out how.

Identifying Relevant Types of Paraphrase

- **Between two discourse types** (technical, lay)
- **Based on 2 hypotheses :**
  - **Technical texts** use more **nominalizations** whereas **lay texts** prefer **verbal constructions (1)**
  - **Technical texts** use more **neo-classical compounds** whereas **lay texts** prefer **modern-language equivalents (2)**
- **Two types of paraphrases :**
  - **nominalization paraphrases**

    treatment of the affection ~ the disease is treated
  - **paraphrases of neo-classical compounds**

    glycemia ~ level of blood sugar

Nominalization Paraphrases

- **Pairs of noun/verb**
  - Deverbal nouns paired with their corresponding verbs

    treatment ~ treat
  - Source : WordNet
- **Paraphrasing rules** built around those pairs
  - Lexico-syntactic patterns developed for French (corpus study)
  - Directly **adapted to English** through simple **syntactic changes**

N1 Prep **N2** ⟶ **N2 V1**

N1 Prep **N2 Adj3** ⟶ **V1 N2 Adj3**

N2 N1 ⟶ **N2 V1**

N1 Prep **Adj3 N2** ⟶ **V1 Adj3 N2**

wound infection ⟶ ***wound is infected***

treatment of infected wounds ⟶ ***treats infected wounds***

Paraphrases of Neo-Classical Compounds

- **Morpho-semantic analysis**
  - DériF analyzer *(Namer 2009; Deléger et al. 2009)*
  - **Decomposition** of neo-classical compounds
  - Acquisition of **modern-language equivalents** for each component

  gastritis *= gastr + itis =* ***stomach*** *+* ***inflammation***
- **Paraphrasing rules** based on this analysis
  - Also **adapted** from a French version through **syntactic changes**

C ⟶ (N Prep) **C1** W{0,4} **C2**

C ⟶ **C1** W{0,4} **C2** (N)

glycemia ⟶ ***blood sugar level***

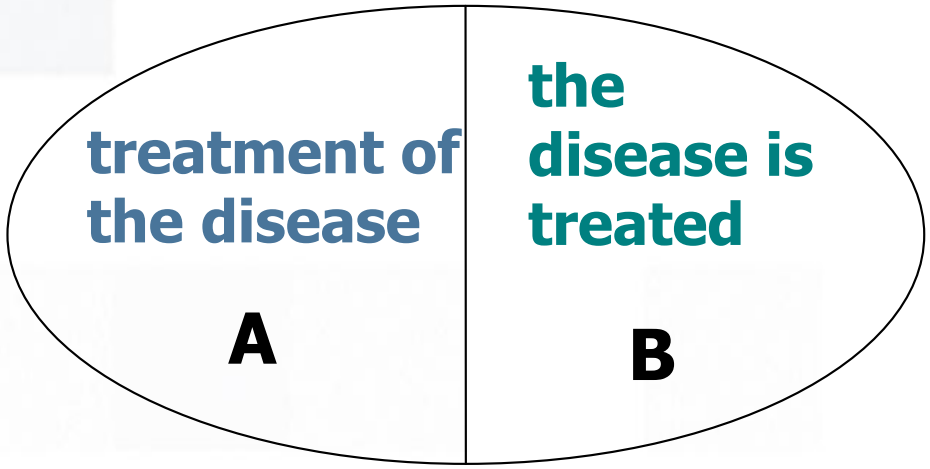
Evaluation

- **Quality of the paraphrases**
  - Measure **precision**
  - 500 sample for nominalization paraphrases
  - All paraphrases of neo-classical compounds

Evaluation

- **Coherence of the initial hypotheses :**
  - Comparison between the technical and lay sides of the corpora
  - By computing a ***preference index*** **I** for nominalizations and for neo-classical compounds

Proportion of nominalizations / nominalizations + verb phrases
Proportion of neo-classical comp. / neo-cl. comp. + modern-language equivalents

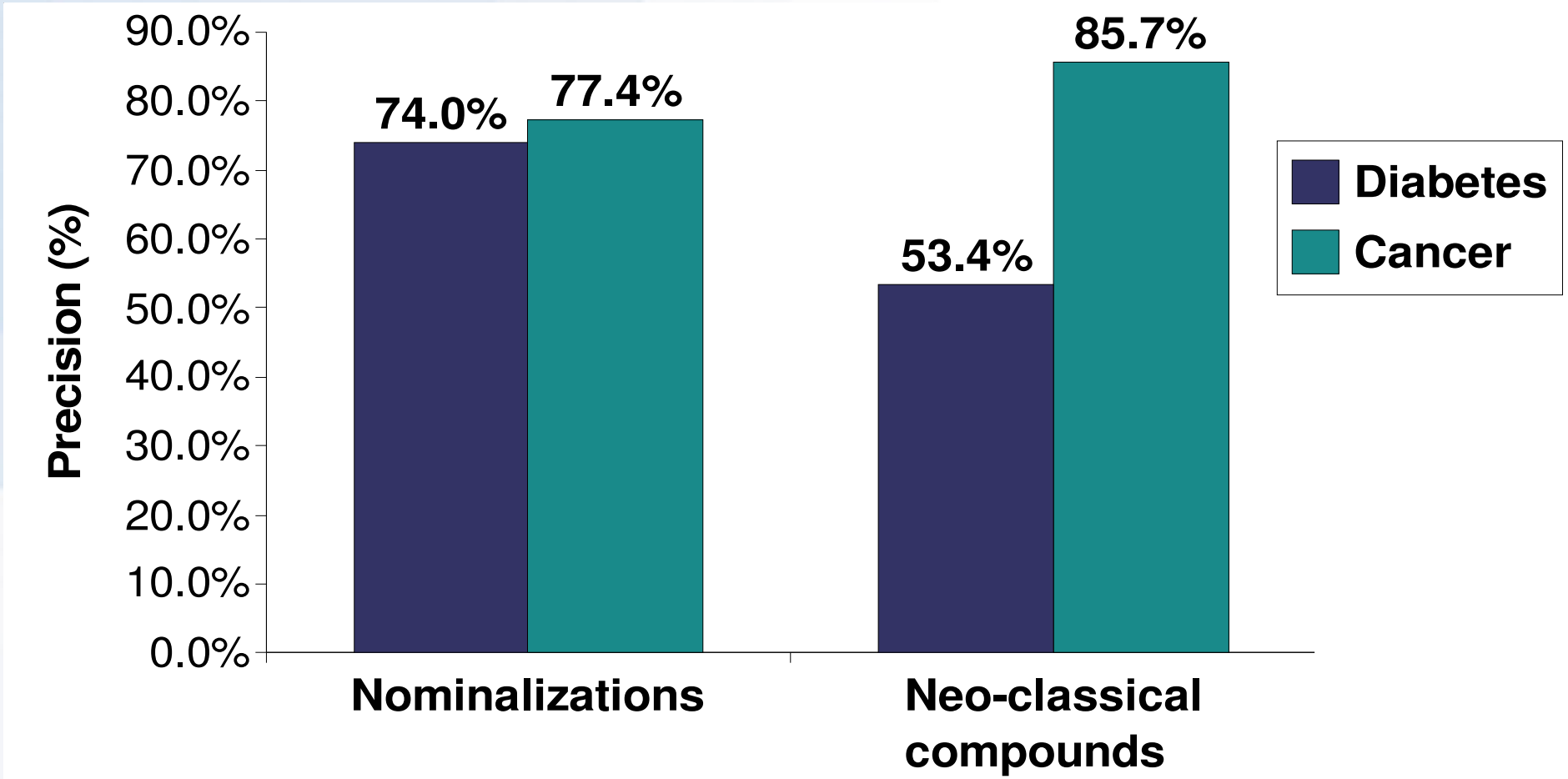


(technical or lay part)

$$I = A / A+B$$

- The index should be **higher for the technical side**

Results : precision

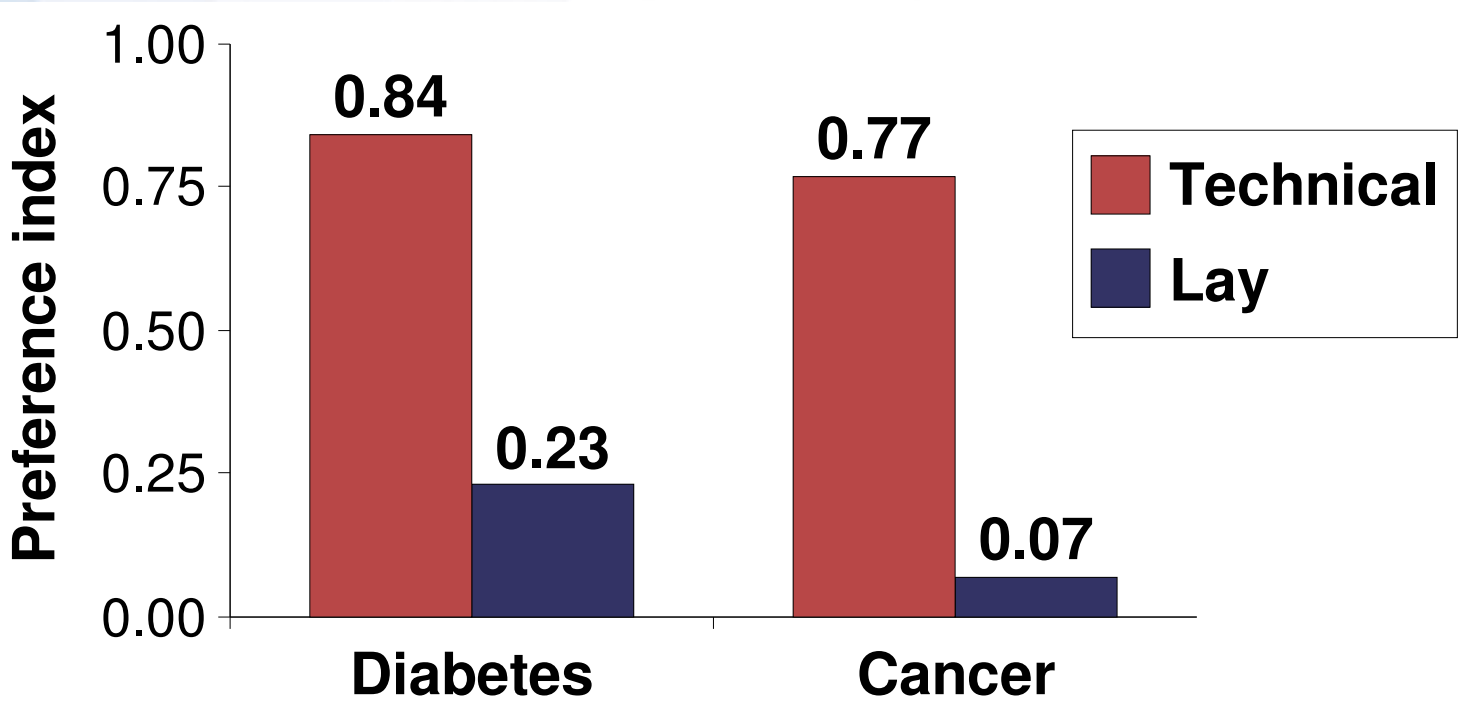


- Nominalizations : good precision for both corpora
- Neo-classical compounds : average and good (but few cases)

~ 14 occurrences

Results :
Coherence with hypothesis 1

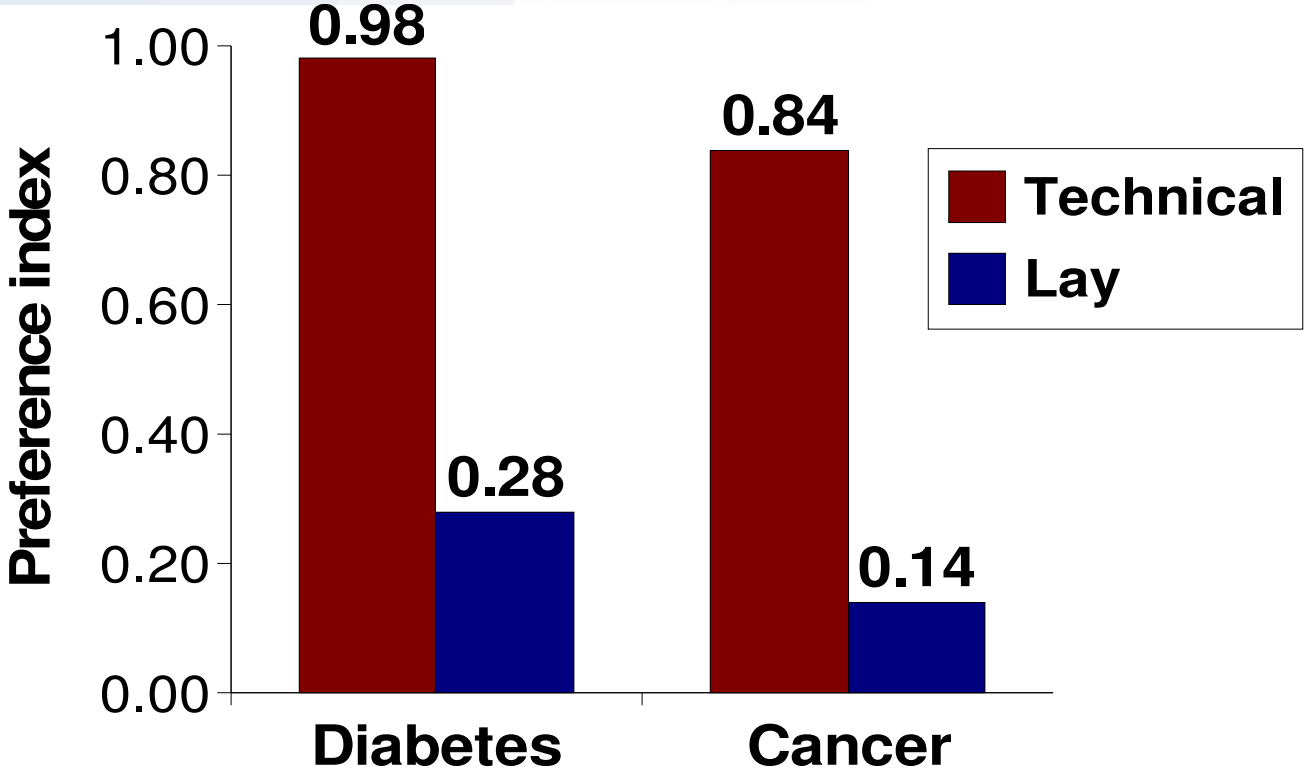
Technical texts use more nominalizations whereas lay texts prefer verbal constructions



- Preference index considerably **higher** in the **technical** side
- Tends to confirm the hypothesis

Results : Coherence with hypothesis 2

Technical texts use more neo-classical compounds whereas lay texts prefer modern-language equivalents



- Preference index considerably **higher** in the **technical** side
- But small number of occurrences : not so conclusive

Results : examples of nominalization paraphrases

Technical	Lay
blood ***replacement***	***replaced*** *blood*
insulin ***absorption***	*insulin is* ***absorbed***
removal *by surgery*	***removed*** *in an operation*
gene ***carriers***	***carry*** *a gene*
practice* ***recommendations*	**exercise is* ***recommended***

Results : examples of paraphrases of neo-classical compounds

Technical	Lay
pancreatitis	***inflammation*** *of the* ***pancreas***
haematuria	***blood*** *in the* ***urine***
erythrocyte	***red*** *blood* ***cell***
amylase	***enzymes*** *that digest* ***starch***
****normoglycemic***	****blood sugar*** *level above* ***normal***

Discussion/Conclusion

- Adaptation of a paraphrase identification method from **French to English**
- **Results are similar** to those obtained for French
  - Paraphrases acquired with a **rather good precision**
  - **Nominalization paraphrases** seem **especially relevant** in the context of the technical vs lay opposition
  - **Less conclusive for neo-classical compounds** given the small number of paraphrases
- **Future work** : new types of paraphrases, larger corpora