



## Introduction to NLP

**Examples of Text** 



## **Understanding a News Story**

#### Brazil crowds attend funeral of late candidate Campos

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday.

They attended a funeral Mass and filled the streets of the city of Recife to follow the passage of his coffin.

Later this week, Mr Campos's Socialist Party is expected to appoint former Environment Minister Marina Silva as a replacement candidate.

Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo.

Investigators are still trying to establish the exact causes of the crash, which killed six other people.

Mr Campos's private plane - a Cessna 560XL - was travelling from Rio de Janeiro to the sea-side resort of Guaruja, near the city of Santos.

President Dilma Rousseff, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.



## **Understanding a News Story**

#### Brazil crowds attend funeral of late candidate Campos

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday.

They attended a funeral Mass and filled the streets of the city of Recife to follow the passage of his coffin.

Later this week, Mr Campos's Socialist Party **is expected** to appoint former Environment Minister Marina Silva as a replacement candidate.

Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo.

Investigators are still trying to establish the exact causes of **the crash**, which killed six other people.

Mr Campos's private plane - a Cessna 560XL - was travelling from Rio de Janeiro to the sea-side resort of Guaruja, near the city of Santos.

**President Dilma Rousseff**, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.

Why did I highlight some of the phrases above?



## Answer to the quiz

- Brazil crowds attend funeral of late candidate Campos
  - Current event
- Mr Campos's jet crashed in bad weather in Santos
  - Background event
- Mr Campos's Socialist Party is expected to appoint...
  - Speculation
- President Dilma Rousseff
  - Property
- They attended a funeral Mass
  - Pronominal reference to an entity in a previous sentence



## **Genres of Text**

- Blogs, emails, press releases, chats, debates, etc.
- Each presents different challenges to NLP







### Induction of Influenza-Specific Mucosal Immunity by an Attenuated Recombinant Sendai Virus

Thuc-vy L. Le1, Elena Mironova2, Dominique Garcin2, Richard W. Compans1\*

- 1 Department of Microbiology and Immunology and Emory Vaccine Center, Emory University School of Medicine, Atlanta, Georgia, United States of America,
- 2 Department of Microbiology and Molecular Medicine, University of Geneva School of Medicine, Geneva, Switzerland

Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. Sendai virus is a member of the Paramyxoviridae family, belongs in the genus respirovirus and shares 60-80% sequence homology to human parainfluenza virus type 1 (HPIV-1) [9,10]. The viral genome consists of a negative sense, non-segmented RNA. Although Sendai virus was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to Sendai virus have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPIV-1. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against HPIV-1 [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and recombinant viruses have been reported to efficiently transduce luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells [16]. These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of parainfluenza virus infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

We have previously described the recovery of a mutant virus GP42-SeV containing a replacement of nucleotides 1–42 of the 3' leader sequence encoding the virus genomic promoter (P<sup>1</sup>), with the corresponding sequence of the 5' trailer sequence encoding the anti-genomic promoter (P<sup>8</sup>) [6]. This mutant exhibits enhanced chronic infection and reduced virus-induced programmed cell death in vitwo [18]. In the present study, we introduced the hemagglutinin (HA) gene from influenza virus A/Puerto Rico/8/1934 (PR/8) into an intergenic sequence of GP42-SeV. This vector, designated GP42-HI, was evaluated for induction of systemic and mucosal antibodies as well as effectiveness of protection from lethal challenge with influenza virus.

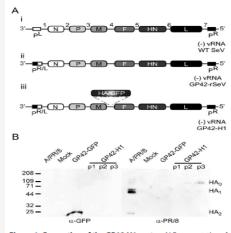


Figure 1. Generation of the GP42-H1 vector. A) Representation of wild-type Sendai virus gene construct (i) illustrating the major viral genes N. P. M. F. HN and L. The left promoter (PL) and right promoter (P<sup>R</sup>) function as the genomic promoter and anti-genomic promoter respectively. Seven gene boundaries that encode the conserved regulatory transcription start and transcription stop sequence are represented numerically. The mutant GP42-SeV (ii) genomic RNA is identical to WT Sendai virus with the exception of 3' PL in which 42 nudeotides of the P<sup>L</sup> were replaced with the corresponding sequence from PR (PR/L). Additional transcription start, stop, poly-adenylation sequences and a unique Mlu I restriction site were introduced into the intergenic region between the Sendai M and Figenes. Using the unique Mlu I restriction site, the GFP or HA gene was inserted (respecting the rule of six) generating the recombinant Sendai GP42-GFP [19] or GP42-H1 vectors (iii). B) Recombinant GP42-H1 virus was cultured in BSR-T7 cells for three passages. Cell free supernatant containing virus suspensions were collected and used to infect CV-1 cells (refer to Materials and Methods), Proteins from GP42-GFP or GP42-H1 infected cell extracts were resolved on SDS-PAGE and screened for GFP (left) or HA (right) expression by western analysis. Mock infected cells and allantoic fluid from PR/8 infected eggs are also shown. doi:10.1371/journal.pone.0018780.g001

Plos ONE

DOI: 10.1371/journal.pone.0018780



Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. Sendai virus is a member of the Paramyxoviridae family, belongs in the genus respirovirus and shares 60–80% sequence homology to human parainfluenza virus type 1 (HPIV-1) [9,10].

The viral genome consists of a negative sense, non-segmented RNA. Although Sendai virus was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to Sendai virus have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPIV-1. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against HPIV-1 [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and recombinant viruses have been reported to efficiently transduce luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells [16].

These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of parainfluenza virus infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

Named entities + variants (human parainfluenza virus type, HPIV-1) Speculation (reported, suggesting) Species (human) Cell types (nasal epithelial cells ) Facts References



### **Medical Records**

```
TITLE: PC ACUTE CARE VISIT
DATE OF NOTE: FEB 04, 2000@11:18
                                     ENTRY DATE: FEB 04, 2000011:20
                                   EXP COSIGNER:
      AUTHOR:
     URGENCY:
                                         STATUS: COMPLETED
Chief Complaint: Patient notes 1 month history of blurred vision and
frequent urination
HISTORY OF PRESENT ILLNESS:
DEMO.FATHER is a 44 year-old MALE who presents complaining of blurred
vision for the past 1 month. He finds it is difficult for him to read
clearly and is even effecting his driving. He also notes that he has be
getting up to the bathroom frequently, esp. at night. He now routinely
to urinate 3-4 times a night. He is not aware of any particular weight
loss, but does feel thirsty much of the time.
PAST MEDICAL HISTORY:
Illnesses: Hypertension
Surgeries: None
Allergies: PENICILLINS
Medications:
   HYDROCHLOROTHIAZIDE 25MG TAB** Qty: 45
       for 90 days Sig: TAKE ONE-HALF TABLET Refills: 0
       MOUTH EVERY MORNING THE BLOOD PRESSURE
2) METOPROLOL 25MG XL TAB Qty: 90 for 90
                                               ACTIVE
       days Sig: TAKE ONE TABLET MOUTH QDAY
                                              Refills: 0
       FOR THE HEART
FAMILY HISTORY:
Diabetes Father, Sibling, Grandparent
```



### **Medical Records**

```
TITLE: PC ACUTE CARE VISIT
DATE OF NOTE: FEB 04, 2000@11:18
                                     ENTRY DATE: FEB 04, 2000011:20
                                   EXP COSIGNER:
      AUTHOR:
     URGENCY:
                                         STATUS: COMPLETED
Chief Complaint: Patient notes 1 month history of blurred vision and
frequent urination
HISTORY OF PRESENT ILLNESS:
DEMO.FATHER is a 44 year-old MALE who presents complaining of blurred
vision for the past 1 month. He finds it is difficult for him to read
clearly and is even effecting his driving. He also notes that he has be
getting up to the bathroom frequently, esp. at night. He now routinely
to urinate 3-4 times a night. He is not aware of any particular weight
loss, but does feel thirsty much of the time.
PAST MEDICAL HISTORY:
Illnesses: Hypertension
Surgeries: None
Allergies: PENICILLINS
Medications:
   HYDROCHLOROTHIAZIDE 25MG TAB** Qty: 45
       for 90 days Sig: TAKE ONE-HALF TABLET Refills: 0
       MOUTH EVERY MORNING THE BLOOD PRESSURE
2) METOPROLOL 25MG XL TAB Qty: 90 for 90
                                               ACTIVE
       days Sig: TAKE ONE TABLET MOUTH QDAY
                                              Refills: 0
       FOR THE HEART
FAMILY HISTORY:
Diabetes Father, Sibling, Grandparent
```



## **Literary Texts**

- Project Gutenberg (http://www.gutenberg.org/browse/scores/top)
- A team of horses passed from Finglas with toiling plodding tread, dragging through the funereal silence a creaking waggon on which lay a granite block. The waggoner marching at their head saluted.
  - Ulysses <a href="http://www.gutenberg.org/files/4300/4300-h/4300-h.htm">http://www.gutenberg.org/files/4300/4300-h/4300-h.htm</a>
- There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question.
  - Jane Eyre http://www.gutenberg.org/files/1260/1260-h/1260-h.htm
- Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife. Their house was small, for the lumber to build it had to be carried by wagon many miles. There were four walls, a floor and a roof, which made one room; and this room contained a rusty looking cookstove, a cupboard for the dishes, a table, three or four chairs, and the beds. Uncle Henry and Aunt Em had a big bed in one corner, and Dorothy a little bed in another corner. There was no garret at all, and no cellar—except a small hole dug in the ground, called a cyclone cellar, where the family could go in case one of those great whirlwinds arose, mighty enough to crush any building in its path. It was reached by a trap door in the middle of the floor, from which a ladder led down into the small, dark hole.
  - The Wizard of Oz http://www.gutenberg.org/files/55/55-h/55-h.htm





## A Really Long Literary Sentence

#### Try parsing this

- "Bloat is one of the co-tenants of the place, a maisonette erected last century, not far from the Chelsea Embankment, by Corydon Throsp, an acquaintance of the Rossettis' who wore hair smocks and liked to cultivate pharmaceutical plants up on the roof (a tradition young Osbie Feel has lately revived), a few of them hardy enough to survive fogs and frosts, but most returning, as fragments of peculiar alkaloids, to rooftop earth, along with manure from a trio of prize Wessex Saddleback sows quartered there by Throsp's successor, and dead leaves off many decorative trees transplanted to the roof by later tenants, and the odd unstomachable meal thrown or vomited there by this or that sensitive epicurean-all got scumbled together, eventually, by the knives of the seasons, to an impasto, feet thick, of unbelievable black topsoil in which anything could grow, not the least being bananas."
- Do you know the source?



# Quiz Answer

- "Gravity's Rainbow" (by Thomas Pynchon), known for its use of very arcane words and complicated sentence (and plot) structure.
- Another such work is "Finnegans Wake" by James Joyce.
- Poetry is even more difficult.

