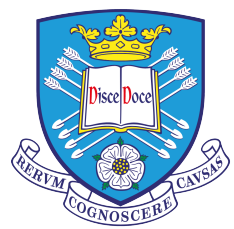
# Learning Simplifications for Specific Target Audiences

Carolina Scarton and Lucia Specia

{c.scarton, l.specia}@sheffield.ac.uk




ACL 2018, Melbourne, Australia

# Text Simplification

If the trend continues, **the researchers say,** some of the rarer amphibians could **disappear in as few as six years** from roughly half the sites where **they're** now found**, while the more common** species could see similar declines in 26 years.

If the trend continues, some of the rarer amphibians could **be gone** from roughly half the sites where **they are** now found **in as few as six years. More common** species could see similar declines in 26 years.

# Text Simplification

If the trend continues, **the researchers say,** some of the rarer amphibians could **disappear** **in as few as six years** from roughly half the sites where **they're** now found**,** **while the more common** species could see similar declines in 26 years.

If the trend continues, some of the rarer amphibians could **be gone** from roughly half the sites where **they are** now found **in as few as six years.** **More common** species could see similar declines in 26 years.

- For a **specific target audience**, e.g. non-native speakers

# Text Simplification

If the trend continues, **the researchers say,** some of the rarer amphibians could **disappear in as few as six years** from roughly half the sites where **they're** now found**, while the more common** species could see similar declines in 26 years.

If the trend continues, some of the rarer amphibians could **be gone** from roughly half the sites where **they are** now found **in as few as six years. More common** species could see similar declines in 26 years.

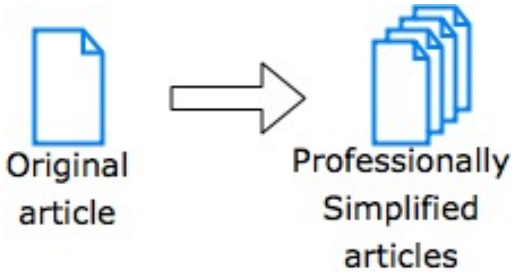
- For a **specific target audience**, e.g. non-native speakers
- For **improving NLP tasks**, e.g. MT

# Newsela Corpus

- Wikipedia – Simple Wikipedia (W–SW)
  - rather **small**
  - not **professionally simplified**
  - no defined **target audience**

# Newsela Corpus

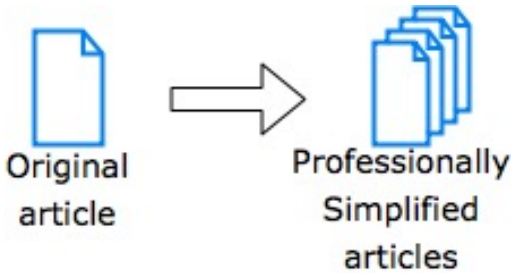
- Wikipedia – Simple Wikipedia (W–SW)
  - rather **small**
  - not **professionally simplified**
  - no defined **target audience**



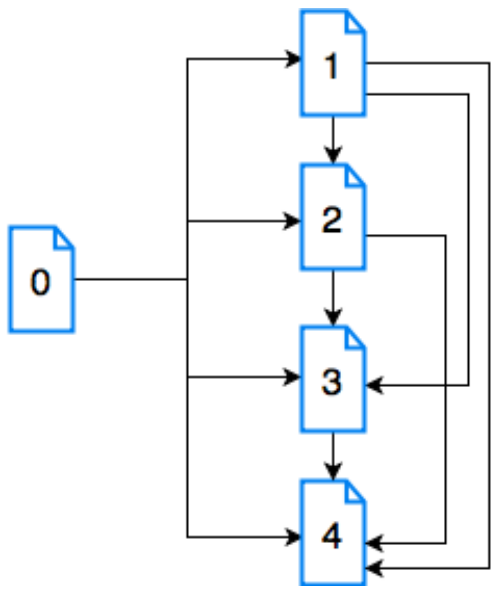

- **Newsela** (version 2016-01-29.1)
  - simplified versions **target different grade levels in the US**
  - professionally simplified

# Newsela Corpus

- Wikipedia – Simple Wikipedia (W–SW)
  - rather **small**
  - not **professionally simplified**
  - no defined **target audience**



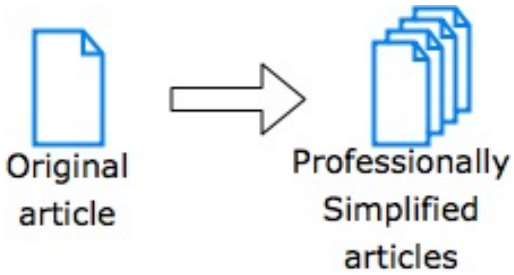

- **Newsela** (version 2016-01-29.1)
  - simplified versions **target different grade levels in the US**
  - professionally simplified



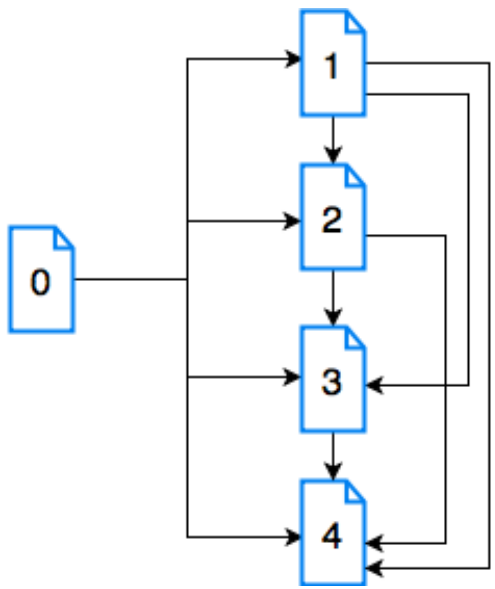

- **Automatic** sentence-level alignments
  - **Identical** (146,251)
  - **Many-to-one (merge)** (24,661)
  - **One-to-many (split)** (121,582)
  - **Elaboration** (258,150)

# Newsela Corpus

- Wikipedia – Simple Wikipedia (W–SW)
  - rather **small**
  - not **professionally simplified**
  - no defined **target audience**




- **Newsela** (version 2016-01-29.1)
  - simplified versions **target different grade levels in the US**
  - professionally simplified




- **Automatic** sentence-level alignments
  - **Identical** (146,251)
  - **Many-to-one (merge)** (24,661)
  - **One-to-many (split)** (121,582)
  - **Elaboration** (258,150)

- **Newsela:** ≈550K sentences pairs (≈ 280K W-SW)

# Sequence-to-Sequence TS

- Sequence-to-Sequence: **state-of-the-art** for other text-to-text transformation tasks
  - NTS [Nisioi et al., 2017] $\rightarrow$ **state-of-the-art on W–SW**

# Sequence-to-Sequence TS

- Sequence-to-Sequence: **state-of-the-art** for other text-to-text transformation tasks
  - NTS [Nisioi et al., 2017] $\rightarrow$ **state-of-the-art on W–SW**
- Previous work disregards **specificities of different audiences**

# Sequence-to-Sequence TS

- Sequence-to-Sequence: **state-of-the-art** for other text-to-text transformation tasks
  - NTS [Nisioi et al., 2017] $\rightarrow$ **state-of-the-art on W–SW**
- Previous work disregards **specificities of different audiences**
- Google's multilingual NMT approach [Johnson et al., 2017]: **artificial token** to guide the encoder

<2es> How are you? $\rightarrow$ Cómo estás?

# Sequence-to-Sequence TS

- Sequence-to-Sequence: **state-of-the-art** for other text-to-text transformation tasks
  - NTS [Nisioi et al., 2017] $\rightarrow$ **state-of-the-art on W–SW**
- Previous work disregards **specificities of different audiences**
- Google's multilingual NMT approach [Johnson et al., 2017]: **artificial token** to guide the encoder

  $<$**2es**$>$ How are you? $\rightarrow$ Cómo estás?

- **Our approach**: artificial token representing the **grade level of the target sentence**

# TS for Different Grade Levels

**< 2 >** dusty handprints stood out against the rust of the fence near Sasabe.

dusty handprints could be seen on the fence near Sasabe.

# TS for Different Grade Levels

- Advantages:
  - **More adequate simplifications** for audiences with different educational levels
  - Real world scenario $\rightarrow$ grade level is **given by the end-user**
  - Robust for **repetitions of source sentences**

# TS for Different Grade Levels

- Advantages:
  - **More adequate simplifications** for audiences with different educational levels
  - Real world scenario $\rightarrow$ grade level is **given by the end-user**
  - Robust for **repetitions of source sentences**

**< 2 >** dusty handprints **stood out against the rust of** the fence near Sasabe.

dusty handprints **could be seen on** the fence near Sasabe.

# TS for Different Grade Levels

- Advantages:
  - **More adequate simplifications** for audiences with different educational levels
  - Real world scenario → grade level is **given by the end-user**
  - Robust for **repetitions of source sentences**

**< 2 >** dusty handprints **stood out against the rust of** the fence near Sasabe.

dusty handprints **could be seen on** the fence near Sasabe.

**< 4 >** dusty handprints **stood out against the rust of** the fence near Sasabe.

dusty handprints **stood out against the rust of** the fence near Sasabe.

# Simplification Operations Information

- Sentence-level alignments → **coarse-grained operations**
  - Identical, Elaborate, Split, Merge

**< elaboration >** dusty handprints **stood out against the rust of** the fence near Sasabe.

dusty handprints **could be seen on** the fence near Sasabe.

# Simplification Operations Information

- Sentence-level alignments → **coarse-grained operations**
  - Identical, Elaborate, Split, Merge

**< elaboration >** dusty handprints **stood out against the rust of** the fence near Sasabe.

dusty handprints **could be seen on** the fence near Sasabe.

- Problem: **not available at test time**

# Simplification Operations Information

- Sentence-level alignments → **coarse-grained operations**
  - Identical, Elaborate, Split, Merge

**< elaboration >** dusty handprints **stood out against the rust of** the fence near Sasabe.

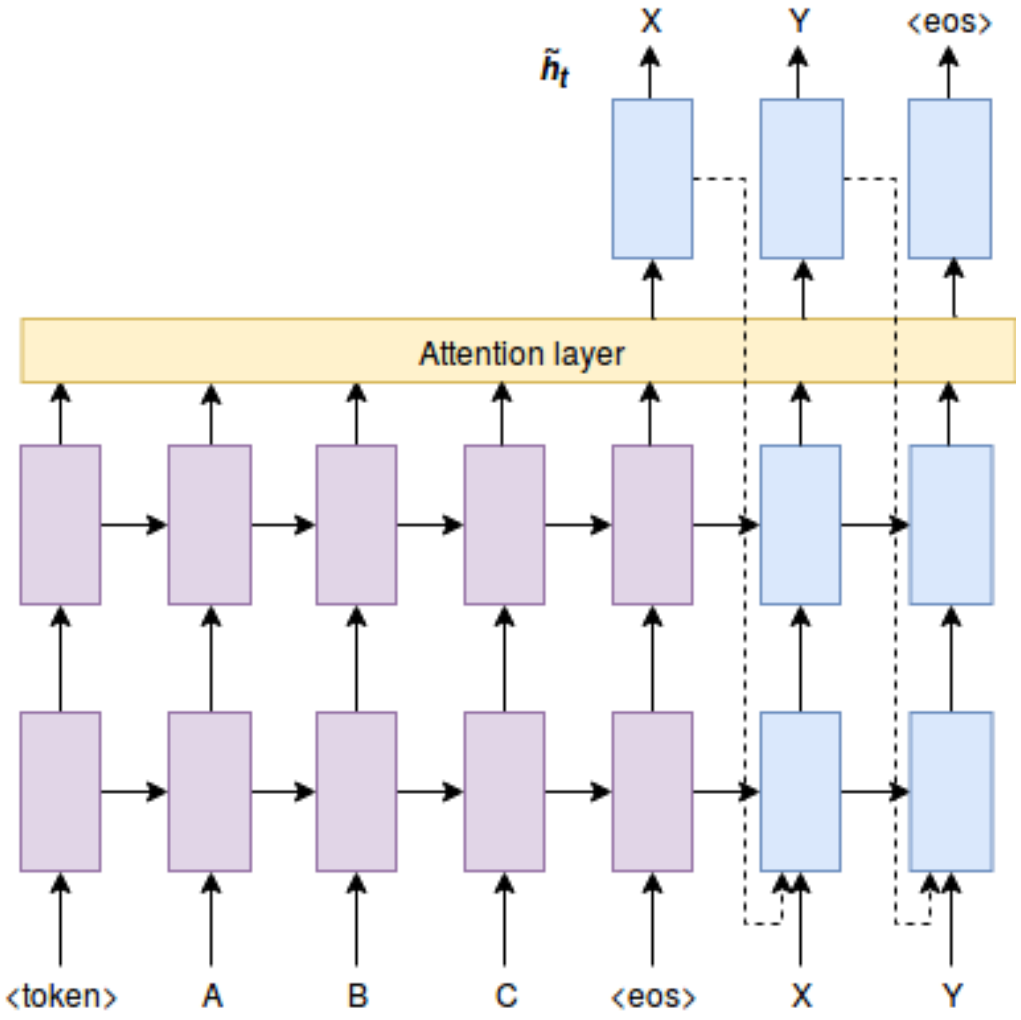
dusty handprints **could be seen on** the fence near Sasabe.

- Problem: **not available at test time**

- **Simplification operations classification**
  - **four-class** classifier → Naive Bayes with **nine features**
  - **Accuracy: 0.51**

# Experiment and results

- NMT approach → **default OpenNMT**



# Experiments and Results

- NTS (w2v): no artificial tokens

# Experiments and Results

- NTS (w2v): no artificial tokens
- **Our models:**
  - s2s (baseline): no artificial tokens
  - s2s+to-grade → **<2>**
  - s2s+operation (pred/gold) → **<elaboration>**
  - s2s+to-grade+operation (pred/gold) → **<2-elaboration>**

# Experiments and Results

- NTS (w2v): no artificial tokens
- **Our models:**
  - s2s (baseline): no artificial tokens
  - s2s+to-grade → <2>
  - s2s+operation (pred/gold) → <elaboration>
  - s2s+to-grade+operation (pred/gold) → <2-elaboration>

| | **BLEU ↑** | **SARI ↑** | **Flesch ↑** |
|---|---|---|---|
| NTS | 61.60 | 33.40 | 79.95 |
| s2s | 61.78 | 33.72 | 79.86 |

# Experiments and Results

- NTS (w2v): no artificial tokens
- **Our models:**
  - s2s (baseline): no artificial tokens
  - s2s+to-grade → <2>
  - s2s+operation (pred/gold) → <elaboration>
  - s2s+to-grade+operation (pred/gold) → <2-elaboration>

| | **BLEU ↑** | **SARI ↑** | **Flesch ↑** |
|---|---|---|---|
| NTS | 61.60 | 33.40 | 79.95 |
| s2s | 61.78 | 33.72 | 79.86 |
| s2s+to-grade | **62.91** | **41.04** | 82.91 |
| s2s+operation (pred) | 59.83 | 37.36 | **84.96** |
| s2s+to-grade+operation (pred) | 61.48 | 40.56 | 83.11 |

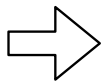
# Experiments and Results

- NTS (w2v): no artificial tokens
- **Our models:**
  - s2s (baseline): no artificial tokens
  - s2s+to-grade → <2>
  - s2s+operation (pred/gold) → <elaboration>
  - s2s+to-grade+operation (pred/gold) → <2-elaboration>

| | **BLEU ↑** | **SARI ↑** | **Flesch ↑** |
|---|---|---|---|
| NTS | 61.60 | 33.40 | 79.95 |
| s2s | 61.78 | 33.72 | 79.86 |
| s2s+to-grade | **62.91** | **41.04** | 82.91 |
| s2s+operation (pred) | 59.83 | 37.36 | **84.96** |
| s2s+to-grade+operation (pred) | 61.48 | 40.56 | 83.11 |
| s2s+operation (gold) | 63.24 | 41.81 | 84.47 |
| s2s+to-grade+operation (gold) | **64.78** | **45.41** | **85.44** |

# Example

Original



We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe.

# Example

| | | |
|---|---|---|
| Original | ⇨ | We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe. |
| s2s Grades 10-7 | ⇨ | We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe. |

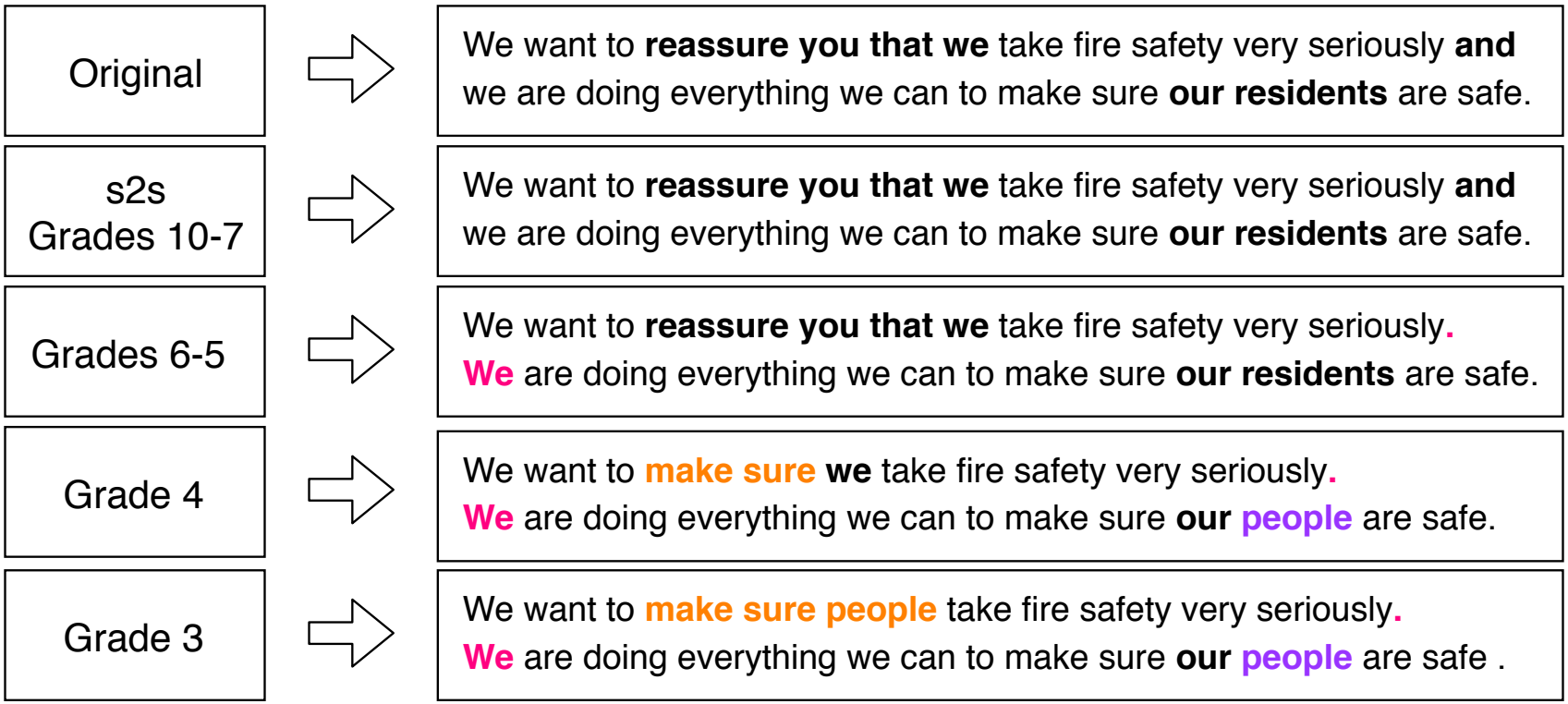
# Example

| | | |
|---|---|---|
| Original | ⇨ | We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe. |
| s2s Grades 10-7 | ⇨ | We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe. |
| Grades 6-5 | ⇨ | We want to **reassure you that we** take fire safety very seriously**.** **We** are doing everything we can to make sure **our residents** are safe. |

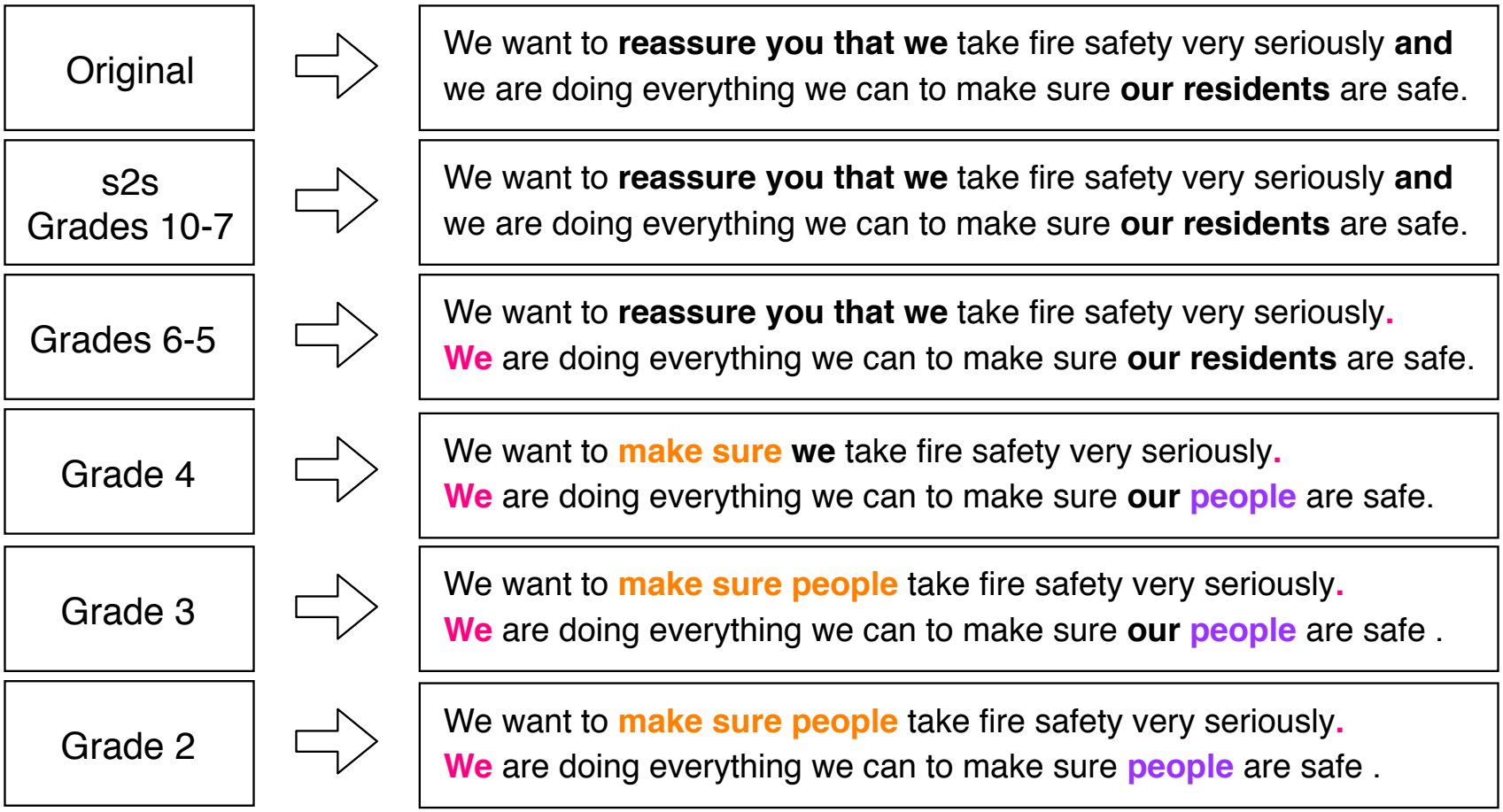
# Example

| | | |
|---|---|---|
| Original | ⇨ | We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe. |
| s2s Grades 10-7 | ⇨ | We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe. |
| Grades 6-5 | ⇨ | We want to **reassure you that we** take fire safety very seriously**.** **We** are doing everything we can to make sure **our residents** are safe. |
| Grade 4 | ⇨ | We want to **make sure** **we** take fire safety very seriously**.** **We** are doing everything we can to make sure **our** **people** are safe. |

# Example


Original
We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe.
s2s Grades 10-7
We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe.
Grades 6-5
We want to **reassure you that we** take fire safety very seriously.
**We** are doing everything we can to make sure **our residents** are safe.
Grade 4
We want to **make sure** **we** take fire safety very seriously.
**We** are doing everything we can to make sure **our** **people** are safe.
Grade 3
We want to **make sure people** take fire safety very seriously.
**We** are doing everything we can to make sure **our** **people** are safe .

# Example


Original
We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe.
s2s Grades 10-7
We want to **reassure you that we** take fire safety very seriously **and** we are doing everything we can to make sure **our residents** are safe.
Grades 6-5
We want to **reassure you that we** take fire safety very seriously.
**We** are doing everything we can to make sure **our residents** are safe.
Grade 4
We want to **make sure we** take fire safety very seriously.
**We** are doing everything we can to make sure **our people** are safe.
Grade 3
We want to **make sure people** take fire safety very seriously.
**We** are doing everything we can to make sure **our people** are safe .
Grade 2
We want to **make sure people** take fire safety very seriously.
**We** are doing everything we can to make sure **people** are safe .

# Zero-shot TS

- **Zero-shot TS** among grade levels
  - Example: **from grade level 12 to grade level 4**
    - **No instances** of 12-to-4 in the training set
    - Other **into 4** levels (e.g. 10-to-4, 6-to-4)

# Zero-shot TS

- **Zero-shot TS** among grade levels
  - Example: **from grade level 12 to grade level 4**
    - **No instances** of 12-to-4 in the training set
    - Other **into 4** levels (e.g. 10-to-4, 6-to-4)

| | **BLEU ↑** | **SARI ↑** | **Flesch ↑** |
|---|---|---|---|
| **12-to-4** | | | |
| s2s | 44.56 | 37.56 | 79.50 |
| s2s+to-grade | 49.43 | 50.76 | 91.04 |
| **s2s+to-grade+zs** | **50.18** | **50.85** | **91.08** |

# Zero-shot TS

- **Zero-shot TS** among grade levels
  - Example: **from grade level 12 to grade level 4**
    - **No instances** of 12-to-4 in the training set
    - Other **into 4** levels (e.g. 10-to-4, 6-to-4)

| | **BLEU ↑** | **SARI ↑** | **Flesch ↑** |
|---|---|---|---|
| **12-to-4** | | | |
| s2s | 44.56 | 37.56 | 79.50 |
| s2s+to-grade | 49.43 | 50.76 | 91.04 |
| **s2s+to-grade+zs** | **50.18** | **50.85** | **91.08** |
| **6-to-5** | | | |
| s2s | **69.71** | **26.47** | 84.74 |
| s2s+to-grade | 69.39 | 26.32 | **87.07** |
| **s2s+to-grade+zs** | 68.78 | 26.23 | 86.80 |

# Conclusions

- TS without target audience $\rightarrow$ **results not ideal**

# Conclusions

- TS without target audience $\rightarrow$ **results not ideal**
- Using a simple **artificial token** with grade level to guide the encoder
  - can **improve the quality of TS**
  - enables **target-audience-oriented** simplifications
  - enables **zero-shot TS**

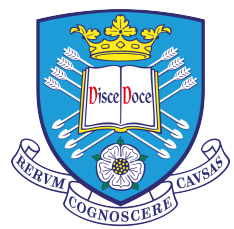
# Conclusions

- TS without target audience $\rightarrow$ **results not ideal**
- Using a simple **artificial token** with grade level to guide the encoder
  - can **improve the quality of TS**
  - enables **target-audience-oriented** simplifications
  - enables **zero-shot TS**
- **Simplification operation** information can help
  - improve classifier for the task
  - explore multi-task learning

# Learning Simplifications for Specific Target Audiences

Carolina Scarton and Lucia Specia

{c.scarton, l.specia}@sheffield.ac.uk




ACL 2018, Melbourne, Australia

# References I


Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017).

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.

TACL, 5:339–351.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017).

Exploring neural text simplification models.

In Proceedings of ACL, pages 85–91.