Evaluating automated translation of LGBTQ+ terms: towards multilingual Homosaurus*

Anna-Maja Kazarian

Shuai Wang

VU Amsterdam

* with support from IHLIA

VU VRIJE UNIVERSITEIT AMSTERDAM

Outline

- Homosaurus
- Motivation and Research questions
- Machine translation
- Survey
- Results
- Potential use cases
- Conclusion
- Future work

Homosaurus

- Vocabulary of 2013 LGBTQ+ terms
- Open data published as linked data using PoolParty
- Not multilingual: Dutch and English vocabularies
- skos:prefLabel, skos:altLabel and rdfs:comment
- Experts are very expensive and can take long time to decide the translations

prefLabel

Donors @en

Donoren @nl
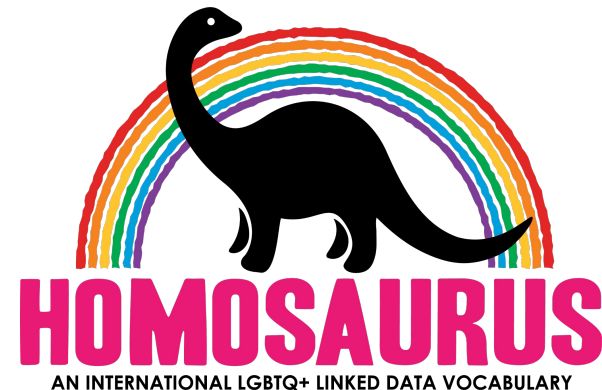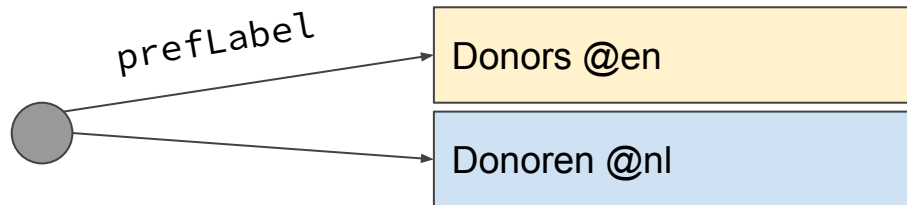
Our project

———

Motivation: why <u>machine translation</u> for LGBTQ+ terms?

- ○ Cheap, fast, convenient
- ○ understand if MTs can translate terms meaningfully

Research questions:

1. How do automated machine translation tools perform on LGBTQ+ terms?
2. How accurate are machine-translation tools on LGBTQ+ terms?
3. Can we use human input to improve the accuracy of automated translation?
4. Does this approach apply for different languages?

Machine Translation

Tools used in this project:

- Google Translate: 133 languages
- Azure: 111 languages
- Reverso: 26 languages
- DeepL: 27 languages
- Libre: 30 languages
- Amazon: 75 languages

Languages:

- Dutch (for RQ1-3)
- German and Spanish (for RQ3-4, popular)
- Polish (for RQ3-4, less popular)

Approach

Interface of the survey

_ _ _

skos:prefLabel   rdfs:comment

Dutch Results: no agreement

What is the best translation of: "Relationship orientation"?

1. Relatie oriëntatie
2. Relatiegerichtheid
3. Oriëntatie op relaties
4. Not sure

Volunteers A & B: Relatie oriëntatie

Volunteers C & D: Relatiegerichtheid

Volunteers E & F: Oriëntatie op relaties

Dutch Results: imperfect translation

———

What is the best translation of: "Surrogacy"?

Google, Azure, Reverso, DeepL, Amazon: **Draagmoederschap**

Libre: **Surrogacy**

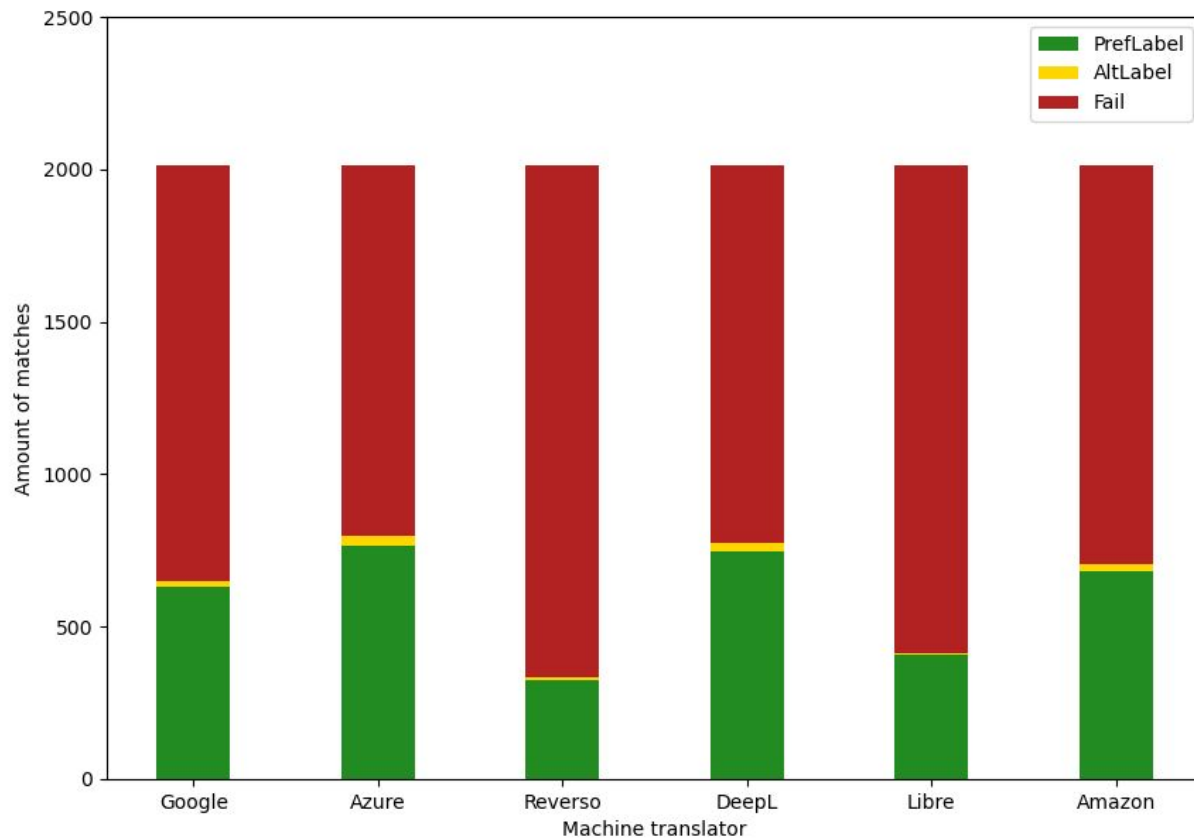Translation in Homosaurus: **Leenouderschap**
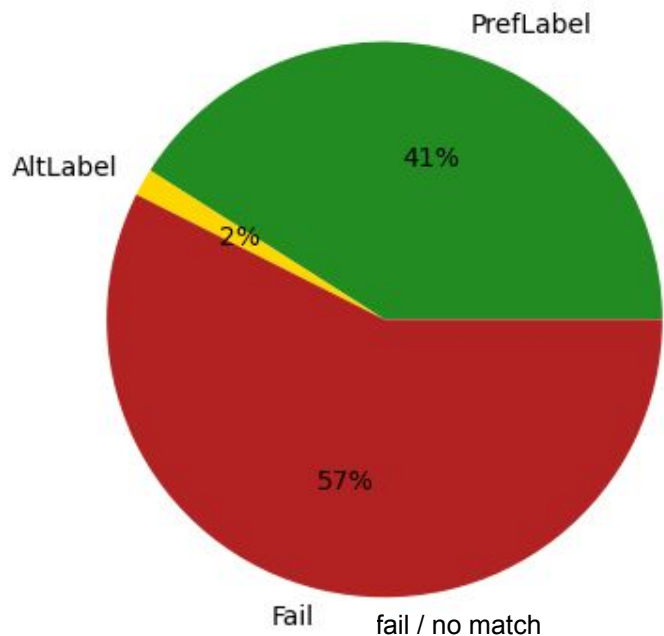
How is this possible?
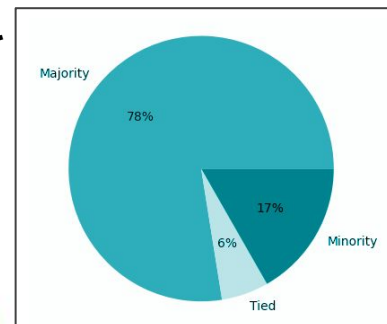
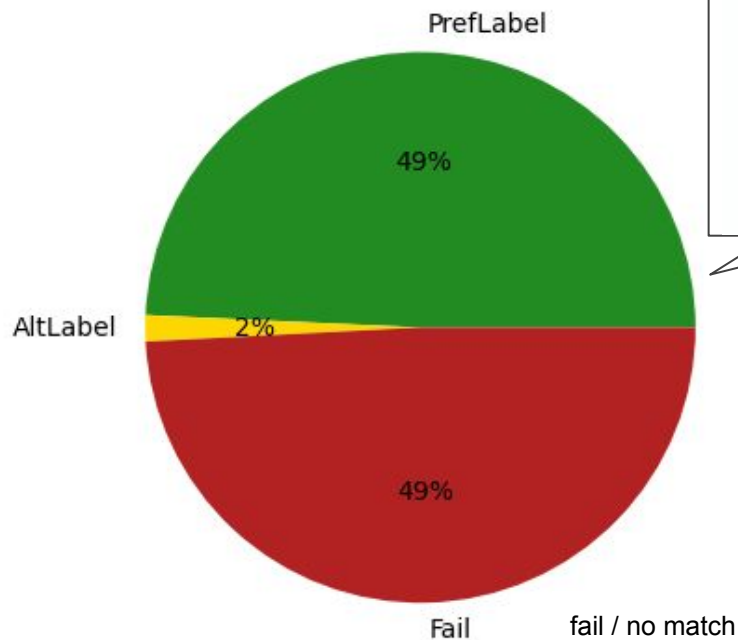# Dutch Results: performance machine translators

– – –

Dutch Results: translation matches with Homosaurus

– – –

Most agreed translations (?%)

At least one translation matches prefLabel or altLabel
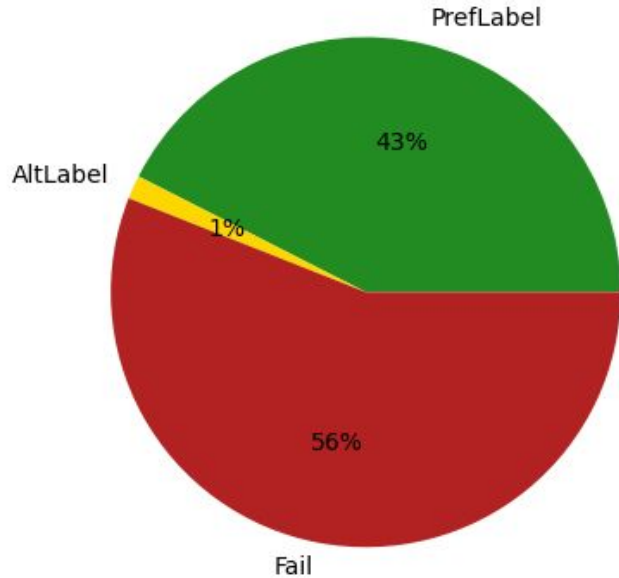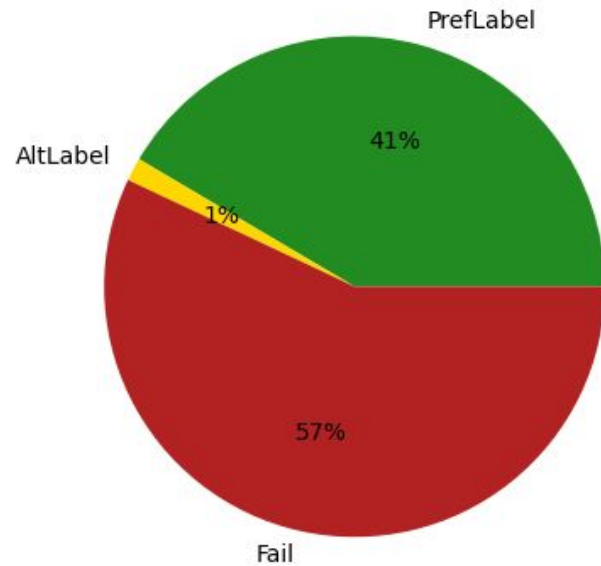
2 Example

———

No match

Dutch Results: comparison MT and human

———

MTs

Volunteers

PrefLabel 43%
AltLabel 1%
Fail 56%

PrefLabel 41%
AltLabel 1%
Fail 57%

**Most agreed translations
(200 terms)**

Towards Multilingual Homosaurus

Multilingual part: no agreement German

———

What is the best translation of: "Gender dysphoria"?

1. Gender-Dysphorie
2. Geschlechtsspezifische Dysphorie
3. Geschlechtskrankheiten
4. Geschlechtsdysphorie
5. Not sure

Volunteer A: Geslechtsdysphorie

Volunteer B: Gender-Dysphorie

Volunteer C: Geschlechtsspezifische Dysphorie

Multilingual part: no agreement Polish

———

What is the best translation of: "Eroticism"?

1. Erotyczność
2. Erotyka
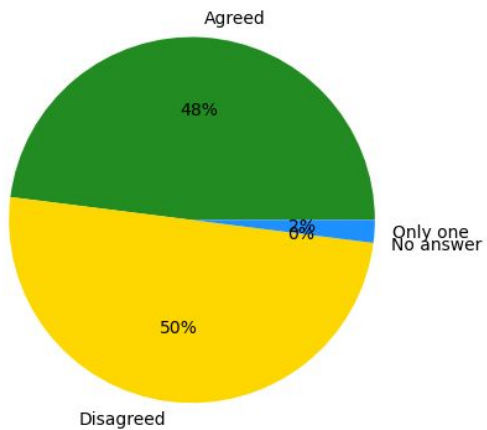3. Erotyzm
4. Not sure

Volunteer A: Erotyzm

Volunteer B: Erotyka

Volunteer C: Erotyczność

Multilingual - Human

- - -

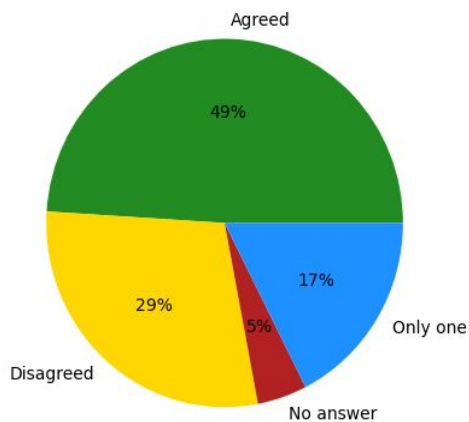Dutch: Agreed 48%, Disagreed 50%, No answer 0%, Only one 2%

Polish: Agreed 55%, Disagreed 22%, No answer 9%, Only one 14%

Spanish: Agreed 49%, Disagreed 29%, No answer 5%, Only one 17%

German: Agreed 43%, Disagreed 32%, No answer 5%, Only one 20%
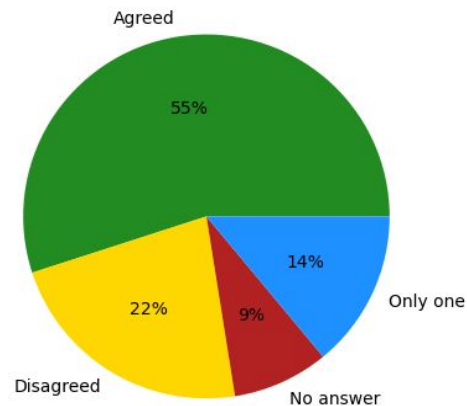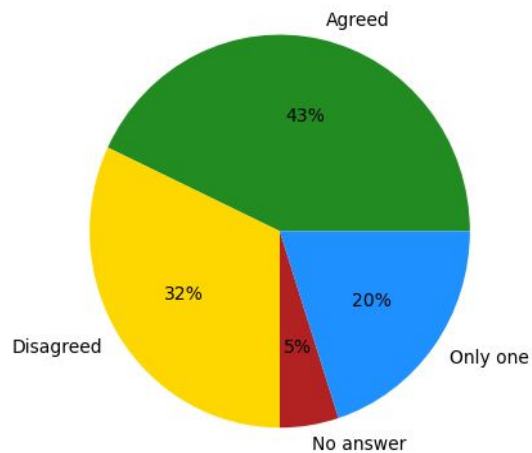
Potential use cases

- Comparing people's attitudes about LGBTQ+ people in social media across languages/country/culture
- Entity linking for libraries, to support searching cross languages
- To support translation of legal documents for lawyers with cases in LGBTQ+

Conclusion

———

- MTs can be useful for experts (to speed up their work)
  - MT terms can suggest terms to experts (49+2% accuracy)
  - MT terms can indicate the term to choose (with 41+2% accuracy)
- The result by majority vote of MTs are highly consistent with that of volunteers
- Without taking the suggestions into account, volunteers are not improving the results of MTs
- Experts are essential for creating multilingual Homosaurus

Future work

‒‒‒

- IHLIA is helping us with translation for evaluation
- The result of this work will be submitted to DHBenelux and the LDK conference.
- Maria & Andrei will continue this work
  - Use weights for translators that perform well
  - Enriching the translation with terms from DBpedia, Wikidata, etc.
  - Translating to more than 4 languages
  - Provide use cases

Thanks for all the support!

___
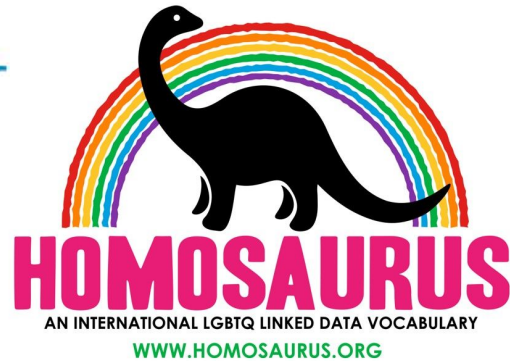
Introduction

___

- LGBTQ+ terms: importance vocabulary
- Homosaurus

Homosaurus

___

- Vocabulary of LGBTQ+ terms
- Goal: enhance discoverability of LGBTQ+ resources
- Linked open data published as linked data in PoolParty
- 2013 terms
- Dutch and English vocabularies
- prefLabel, altLabel and descriptions

○ Homosarus is not multilingual
○ Experts are very expensive and can take long time to decide the translations
○ For many languages, there is no LGBTQ vocabulary yet

poolparty

HOMOSAURUS

AN INTERNATIONAL LGBTQ+ LINKED DATA VOCABULARY
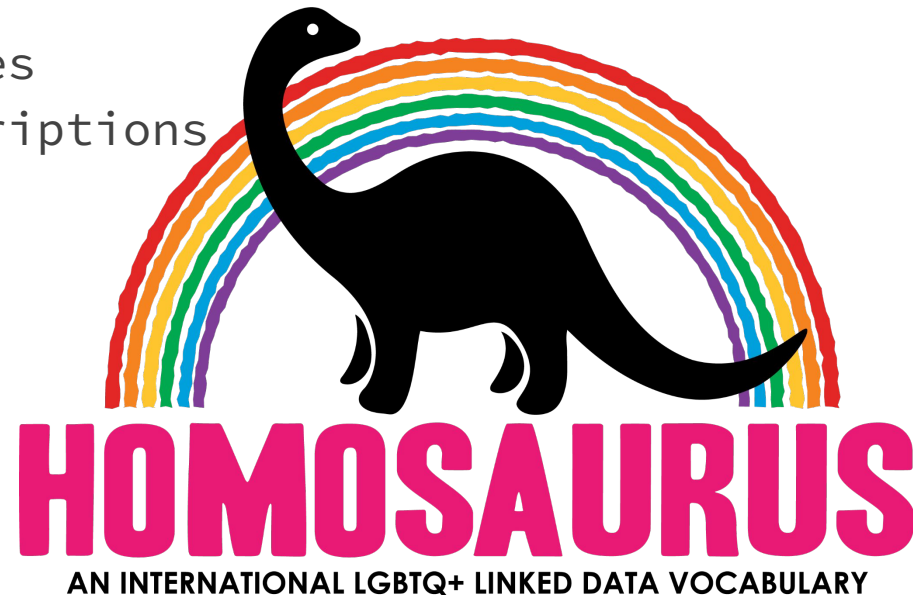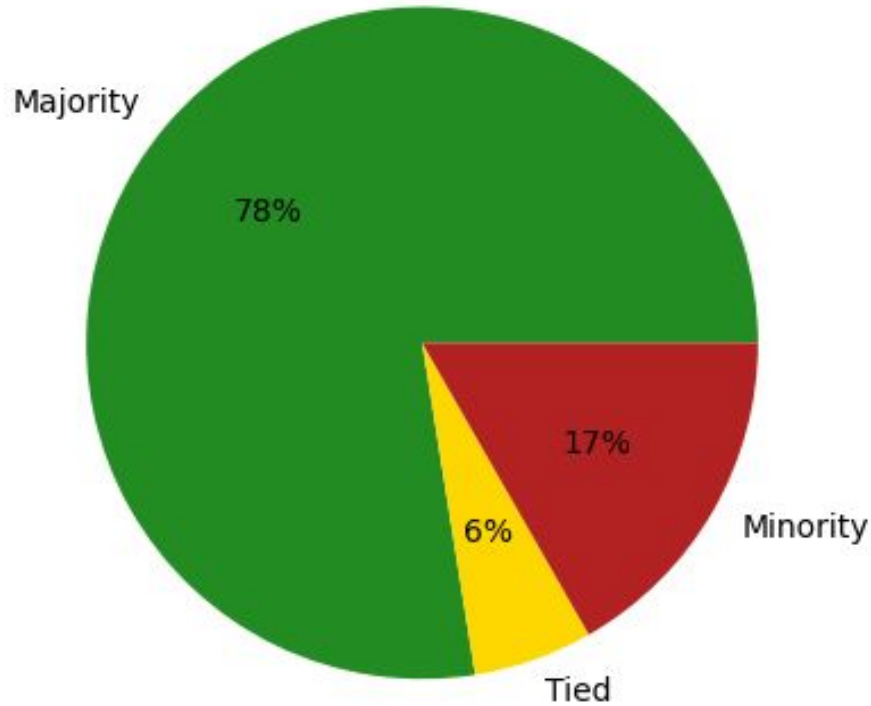
Machine Translation

Approach using machine translators:

1. Translate terms from English to Dutch, German, Polish and Spanish
2. Compare Dutch machine translations to Dutch Homosaurus
3. Compile survey in the 4 different languages

Dutch Results: matches of MTs with prefLabel in Homosaurus
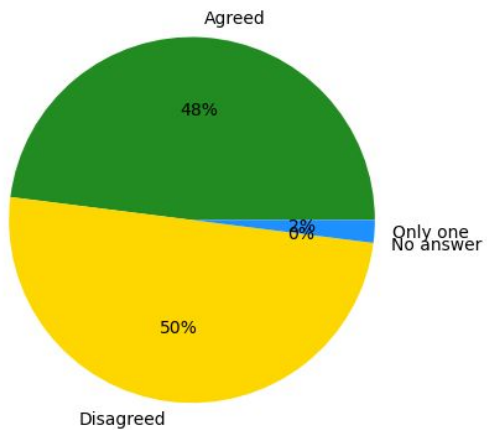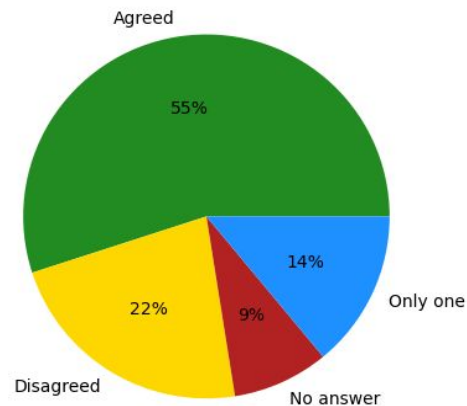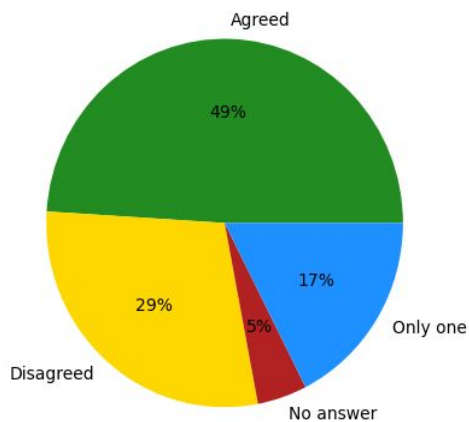
———

Change the color of this plot

Majority 78%

Tied 6%

Minority 17%

Multilingual - MTs

———

Dutch
- Agreed 48%
- Disagreed 50%
- Only one 2%
- No answer 0%

Polish
- Agreed 55%
- Disagreed 22%
- No answer 9%
- Only one 14%

Spanish
- Agreed 49%
- Disagreed 29%
- No answer 5%
- Only one 17%

German
- Agreed 43%
- Disagreed 32%
- No answer 5%
- Only one 20%

Multilingual

___