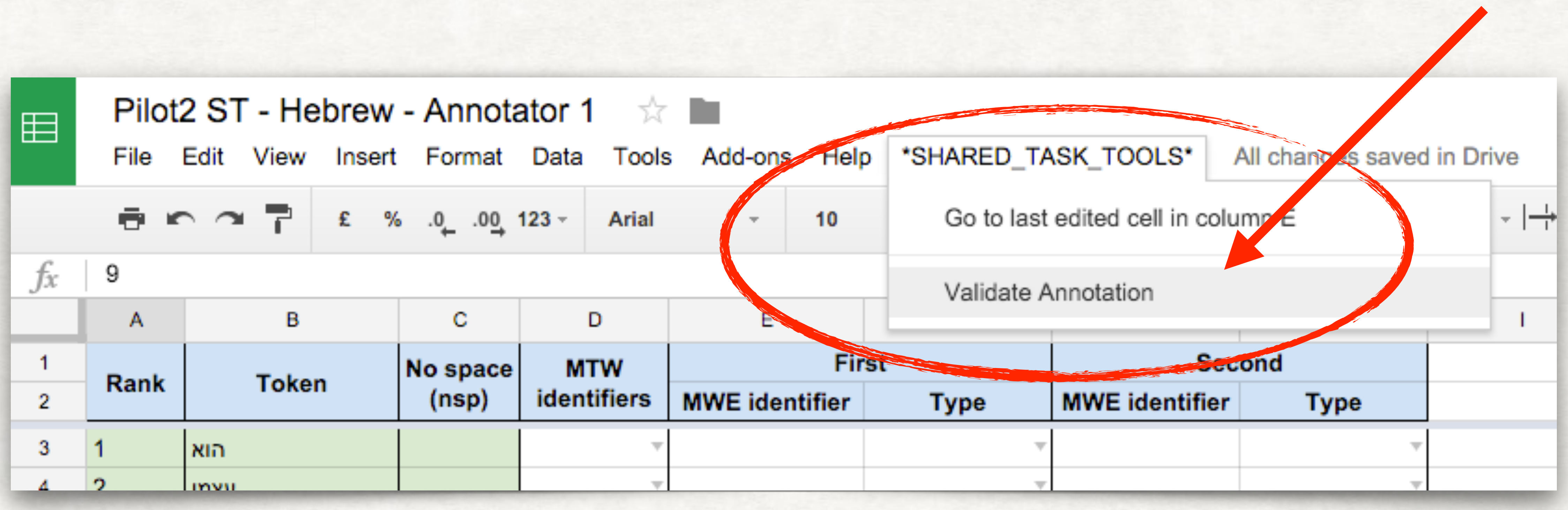
PARSEME

SHARED TASK ANNOTATION TOOLS

Struga, 8 April 2016

PHASE 2

PHASE 2
VALIDATION



	Rank	Token	No space (nsp)	MTW identifiers	First		Second	
					MWE identifier	Type	MWE identifier	Type
3	1	הוא						
4	2							

PHASE 2
VALIDATION

http://tiny.cc/parsemeValidate

↓

ParsemeValidatePhase2.jar

↓

```
java -jar ParsemeValidatePhase2.jar Language
```

PHASE 2
VALIDATION

```
java -jar ParsemeValidatePhase2.jar Swedish

Annotation with double annotate root type or not root
type?! or even using the same number for two different
annotations...
** -- File --> Pilot2 ST - Swedish - Annotator 1.tsv
** -- At line number --> 188
** -- The Token is --> 16   lägga   1  VPC
```

❌

Italian, Hebrew, Romanian, Swedish, Turkish,
English, Portuguese, Maltese, Spanish, Greek

PHASE 2
VALIDATION

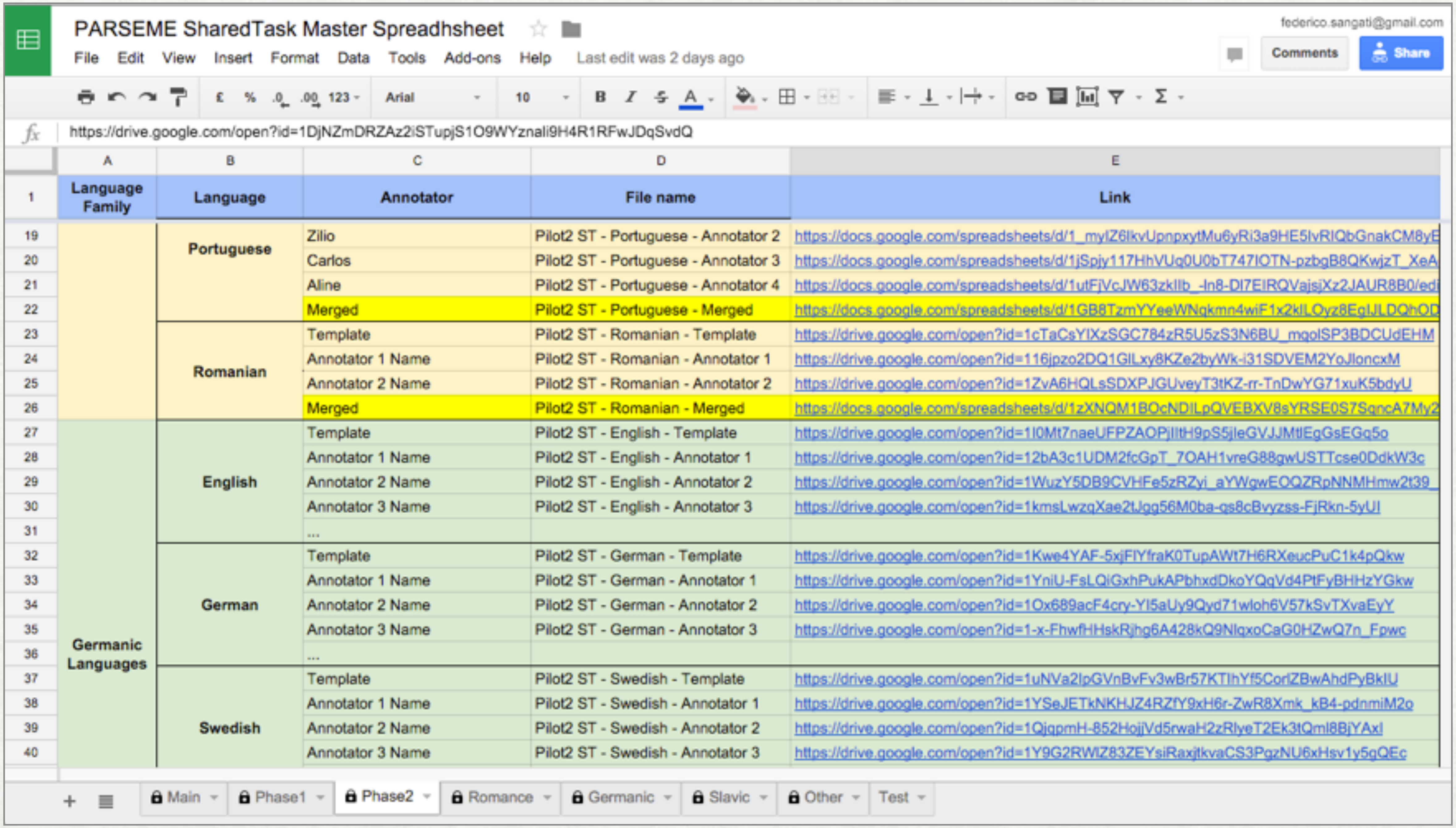
```
java -jar ParsemeValidatePhase2.jar Polish

All the annotations seem to be in good order!
```

✅

Polish, German, French, Hungarian

PHASE 2

MERGING



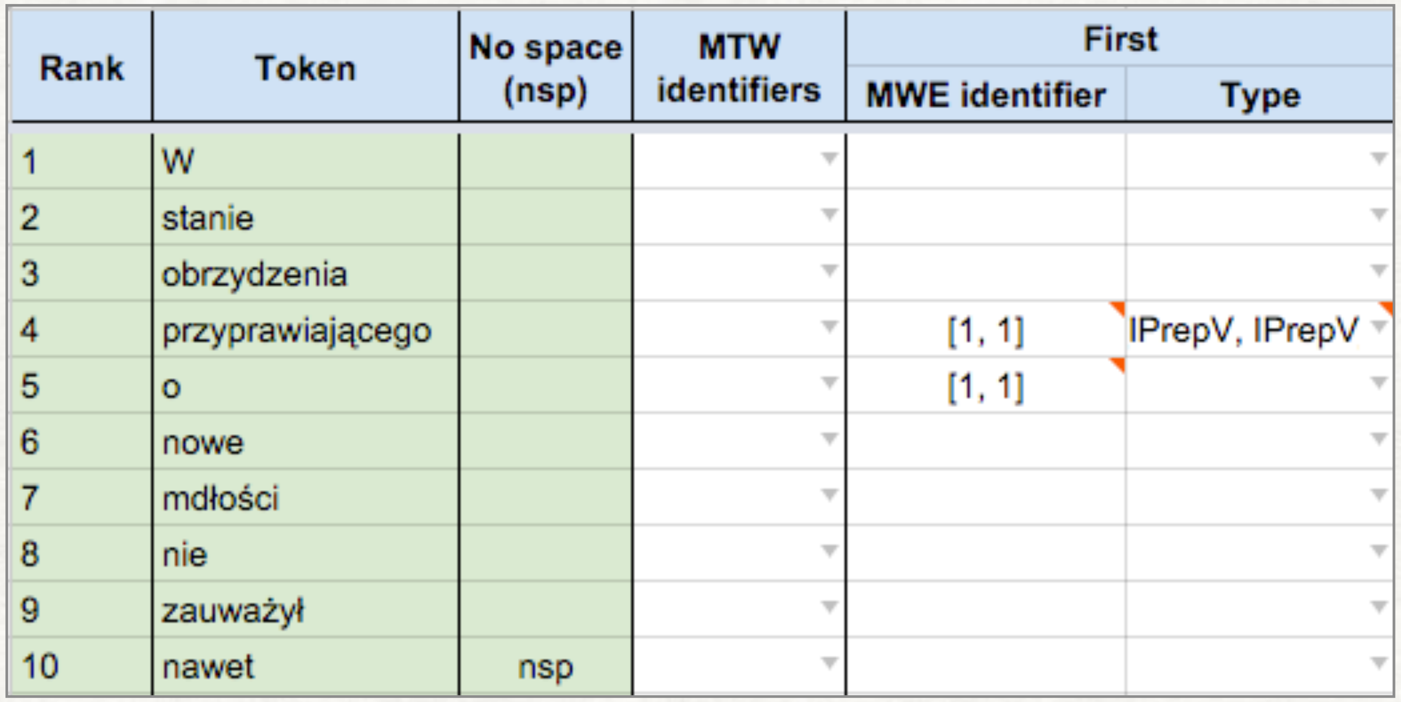
PARSEME SharedTask Master Spreadhsheet

https://drive.google.com/open?id=1DjNZmDRZAz2iSTupjS1O9WYznali9H4R1RFwJDqSvdQ

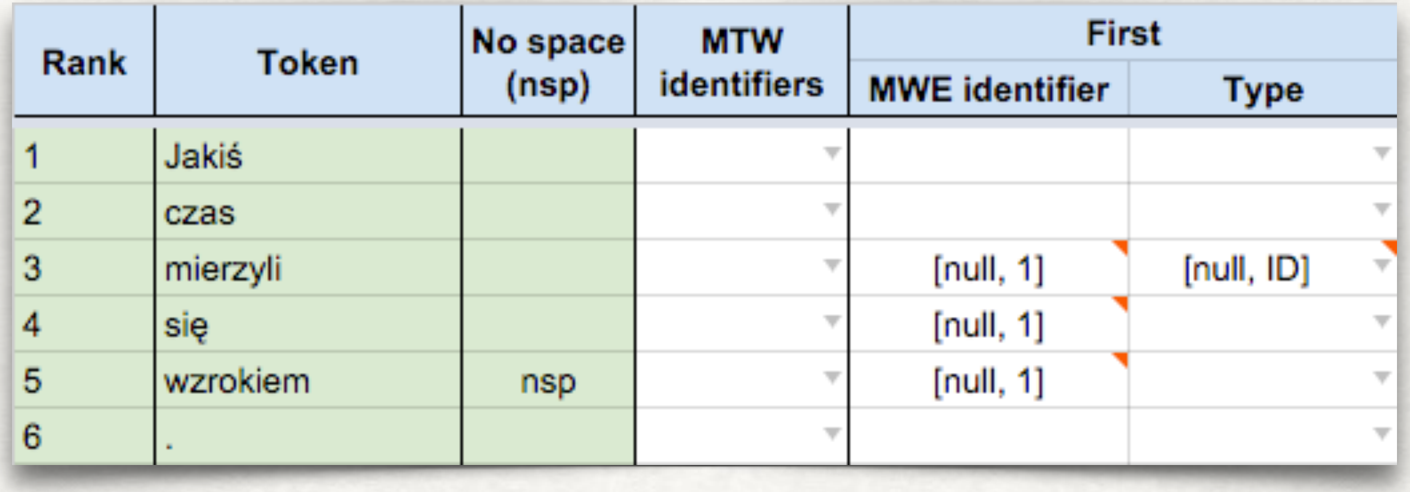
	A	B	C	D	E
1	Language Family	Language	Annotator	File name	Link
19		Portuguese	Zilio	Pilot2 ST - Portuguese - Annotator 2	https://docs.google.com/spreadsheets/d/1_myIZ6IkvUpnpxytMu6yRi3a9HE5IvRIQbGnakCM8yB
20			Carlos	Pilot2 ST - Portuguese - Annotator 3	https://docs.google.com/spreadsheets/d/1jSpjy117HhVUq0U0bT747IOTN-pzbgB8QKwjzT_XeA
21			Aline	Pilot2 ST - Portuguese - Annotator 4	https://docs.google.com/spreadsheets/d/1utFjVcJW63zkIIb_-In8-DI7EIRQVajsjXz2JAUR8B0/edi
22			Merged	Pilot2 ST - Portuguese - Merged	https://docs.google.com/spreadsheets/d/1GB8TzmYYeeWNqkmn4wiF1x2kILOyz8EgIJLDQhOD
23		Romanian	Template	Pilot2 ST - Romanian - Template	https://drive.google.com/open?id=1cTaCsYIXzSGC784zR5U5zS3N6BU_mqolSP3BDCUdEHM
24			Annotator 1 Name	Pilot2 ST - Romanian - Annotator 1	https://drive.google.com/open?id=116jpzo2DQ1GlLxy8KZe2byWk-i31SDVEM2YoJloncxM
25			Annotator 2 Name	Pilot2 ST - Romanian - Annotator 2	https://drive.google.com/open?id=1ZvA6HQLsSDXPJGUveyT3tKZ-rr-TnDwYG71xuK5bdyU
26			Merged	Pilot2 ST - Romanian - Merged	https://docs.google.com/spreadsheets/d/1zXNQM1BOcNDILpQVEBXV8sYRSE0S7SqncA7My2
27	Germanic Languages	English	Template	Pilot2 ST - English - Template	https://drive.google.com/open?id=1I0Mt7naeUFPZAOPjlItH9pS5jleGVJJMtlEgGsEGq5o
28			Annotator 1 Name	Pilot2 ST - English - Annotator 1	https://drive.google.com/open?id=12bA3c1UDM2fcGpT_7OAH1vreG88gwUSTTcse0DdkW3c
29			Annotator 2 Name	Pilot2 ST - English - Annotator 2	https://drive.google.com/open?id=1WuzY5DB9CVHFe5zRZyi_aYWgwEOQZRpNNMHmw2t39_
30			Annotator 3 Name	Pilot2 ST - English - Annotator 3	https://drive.google.com/open?id=1kmsLwzqXae2tJgg56M0ba-qs8cBvyzss-FjRkn-5yUI
31			...		
32		German	Template	Pilot2 ST - German - Template	https://drive.google.com/open?id=1Kwe4YAF-5xjFlYfraK0TupAWt7H6RXeucPuC1k4pQkw
33			Annotator 1 Name	Pilot2 ST - German - Annotator 1	https://drive.google.com/open?id=1YniU-FsLQiGxhPukAPbhxdDkoYQqVd4PtFyBHHzYGkw
34			Annotator 2 Name	Pilot2 ST - German - Annotator 2	https://drive.google.com/open?id=1Ox689acF4cry-YI5aUy9Qyd71wloh6V57kSvTXvaEyY
35			Annotator 3 Name	Pilot2 ST - German - Annotator 3	https://drive.google.com/open?id=1-x-FhwfHHskRjhg6A428kQ9NlqxoCaG0HZwQ7n_Fpwc
36			...		
37		Swedish	Template	Pilot2 ST - Swedish - Template	https://drive.google.com/open?id=1uNVa2IpGVnBvFv3wBr57KTIhYf5CorlZBwAhdPyBkIU
38			Annotator 1 Name	Pilot2 ST - Swedish - Annotator 1	https://drive.google.com/open?id=1YSeJETkNKHJZ4RZfY9xH6r-ZwR8Xmk_kB4-pdnmiM2o
39			Annotator 2 Name	Pilot2 ST - Swedish - Annotator 2	https://drive.google.com/open?id=1QjqpmH-852HojjVd5rwaH2zRlyeT2Ek3tQml8BjYAxl
40			Annotator 3 Name	Pilot2 ST - Swedish - Annotator 3	https://drive.google.com/open?id=1Y9G2RWIZ83ZEYsiRaxjtkvaCS3PgzNU6xHsv1y5gQEc

PHASE 2

MERGING



Rank	Token	No space (nsp)	MTW identifiers	First	
				MWE identifier	Type
1	W				
2	stanie				
3	obrzydzenia				
4	przyprawiającego			[1, 1]	IPrepV, IPrepV
5	o			[1, 1]	
6	nowe				
7	mdłości				
8	nie				
9	zauważył				
10	nawet	nsp			



Rank	Token	No space (nsp)	MTW identifiers	First	
				MWE identifier	Type
1	Jakiś				
2	czas				
3	mierzyli			[null, 1]	[null, ID]
4	się			[null, 1]	
5	wzrokiem	nsp		[null, 1]	
6	.				

ANNOTATION TOOLS

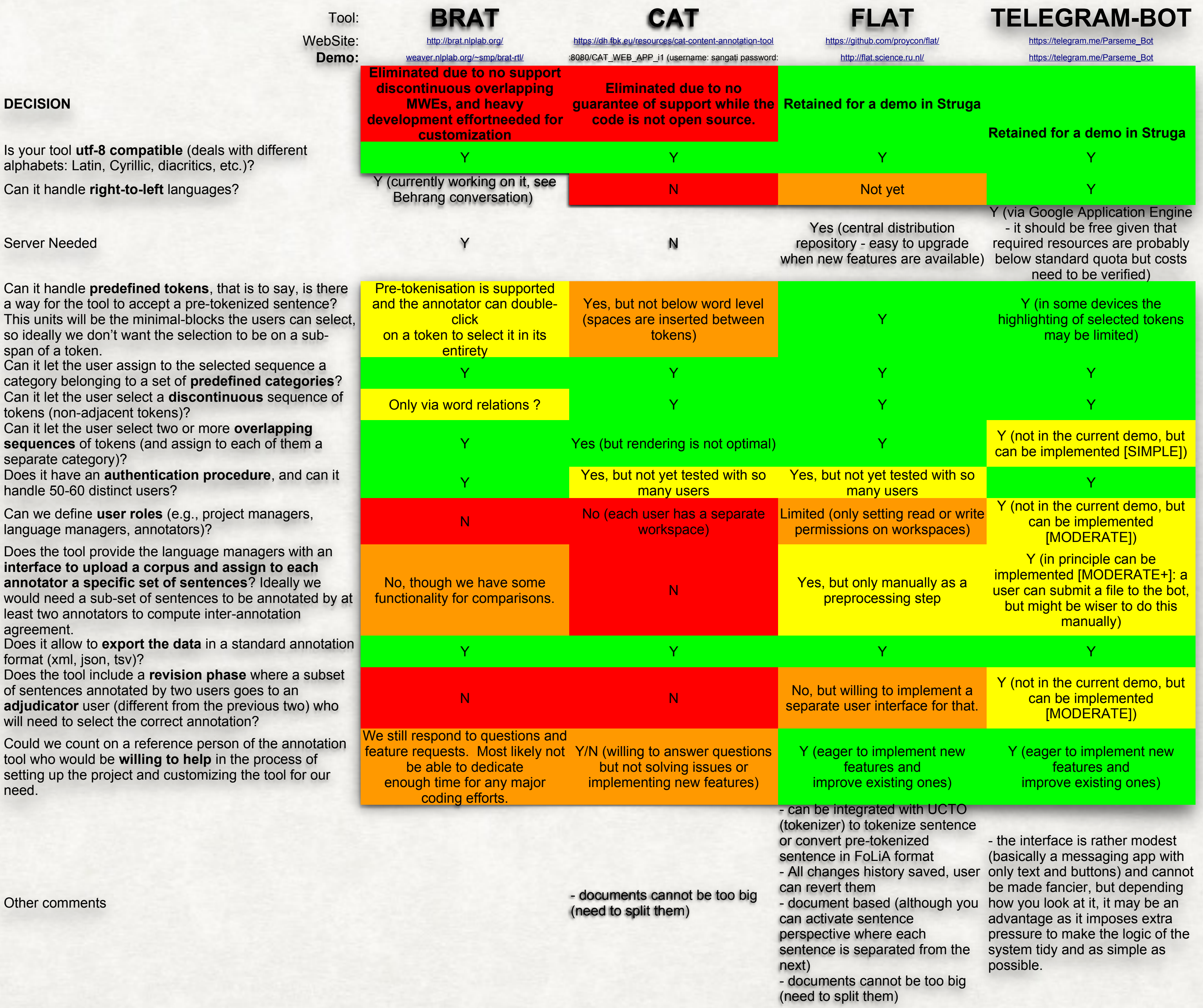
TOOL SELECTION PROCESS

BRAT
CAT
FLAT
CATMA
WebAnno
MAT
Pub annotation
TELEGRAM-BOT

↓

BRAT
CAT
FLAT
TELEGRAM-BOT

TOOL SELECTION PROCESS

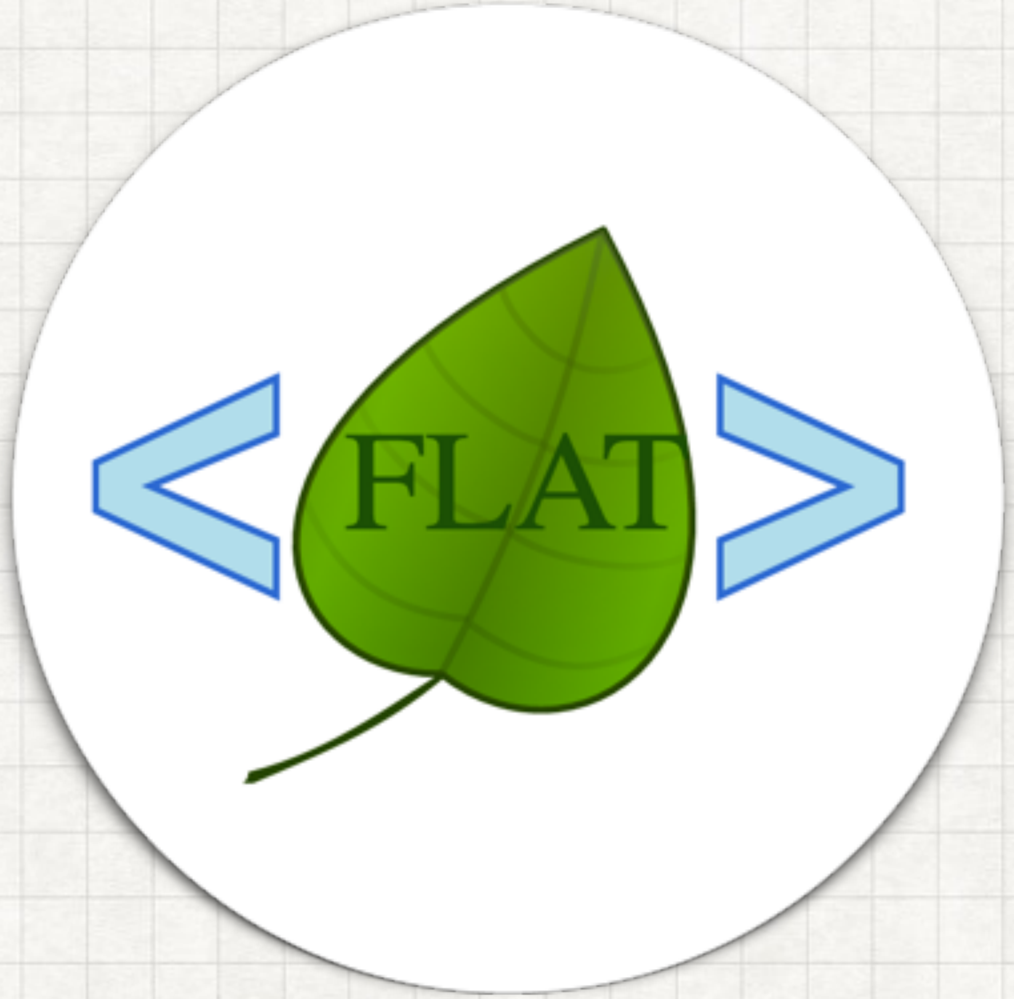


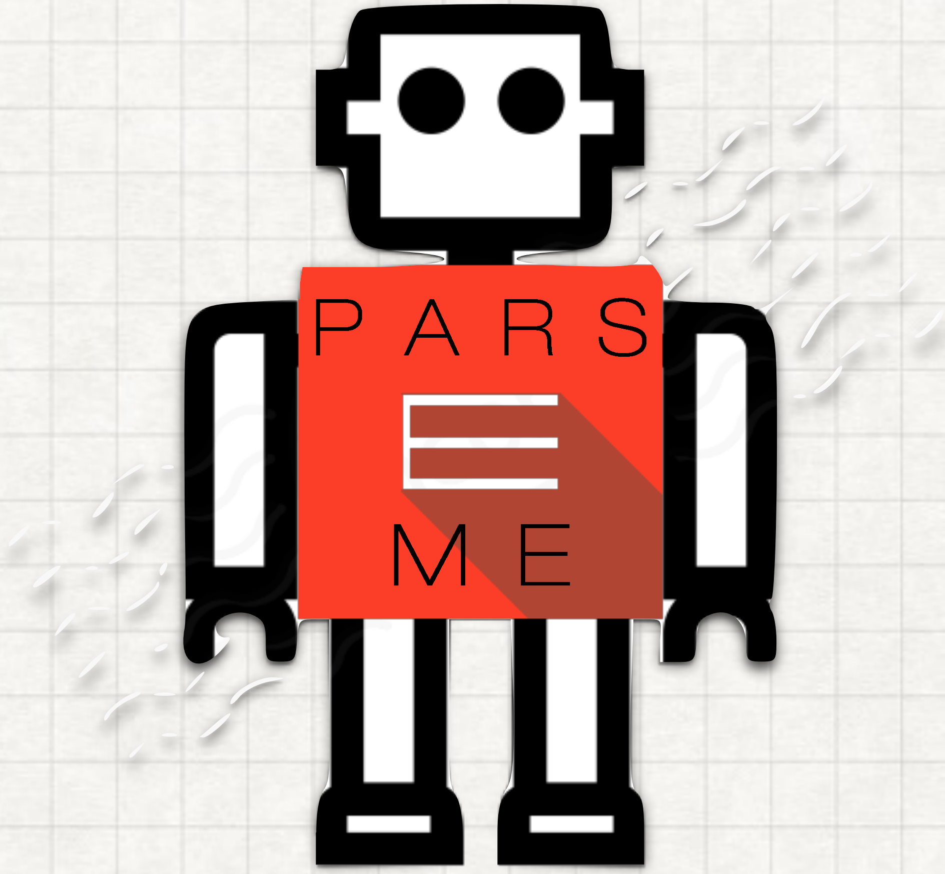
Tool:	BRAT	CAT	FLAT	TELEGRAM-BOT
WebSite:	http://brat.nlplab.org/	https://dh.fbk.eu/resources/cat-content-annotation-tool	https://github.com/proycon/flat/	https://telegram.me/Parseme_Bot
Demo:	weaver.nlplab.org/~smp/brat-rtl/	:8080/CAT_WEB_APP_i1 (username: sangati password:	http://flat.science.ru.nl/	https://telegram.me/Parseme_Bot
DECISION	**Eliminated due to no support discontinuous overlapping MWEs, and heavy development effortneeded for customization**	**Eliminated due to no guarantee of support while the code is not open source.**	**Retained for a demo in Struga**	**Retained for a demo in Struga**
Is your tool **utf-8 compatible** (deals with different alphabets: Latin, Cyrillic, diacritics, etc.)?	Y	Y	Y	Y
Can it handle **right-to-left** languages?	Y (currently working on it, see Behrang conversation)	N	Not yet	Y
Server Needed	Y	N	Yes (central distribution repository - easy to upgrade when new features are available)	Y (via Google Application Engine - it should be free given that required resources are probably below standard quota but costs need to be verified)
Can it handle **predefined tokens**, that is to say, is there a way for the tool to accept a pre-tokenized sentence? This units will be the minimal-blocks the users can select, so ideally we don't want the selection to be on a sub-span of a token.	Pre-tokenisation is supported and the annotator can double-click on a token to select it in its entirety	Yes, but not below word level (spaces are inserted between tokens)	Y	Y (in some devices the highlighting of selected tokens may be limited)
Can it let the user assign to the selected sequence a category belonging to a set of **predefined categories**?	Y	Y	Y	Y
Can it let the user select a **discontinuous** sequence of tokens (non-adjacent tokens)?	Only via word relations ?	Y	Y	Y
Can it let the user select two or more **overlapping sequences** of tokens (and assign to each of them a separate category)?	Y	Yes (but rendering is not optimal)	Y	Y (not in the current demo, but can be implemented [SIMPLE])
Does it have an **authentication procedure**, and can it handle 50-60 distinct users?	Y	Yes, but not yet tested with so many users	Yes, but not yet tested with so many users	Y
Can we define **user roles** (e.g., project managers, language managers, annotators)?	N	No (each user has a separate workspace)	Limited (only setting read or write permissions on workspaces)	Y (not in the current demo, but can be implemented [MODERATE])
Does the tool provide the language managers with an **interface to upload a corpus and assign to each annotator a specific set of sentences**? Ideally we would need a sub-set of sentences to be annotated by at least two annotators to compute inter-annotation agreement.	No, though we have some functionality for comparisons.	N	Yes, but only manually as a preprocessing step	Y (in principle can be implemented [MODERATE+]: a user can submit a file to the bot, but might be wiser to do this manually)
Does it allow to **export the data** in a standard annotation format (xml, json, tsv)?	Y	Y	Y	Y
Does the tool include a **revision phase** where a subset of sentences annotated by two users goes to an **adjudicator** user (different from the previous two) who will need to select the correct annotation?	N	N	No, but willing to implement a separate user interface for that.	Y (not in the current demo, but can be implemented [MODERATE])
Could we count on a reference person of the annotation tool who would be **willing to help** in the process of setting up the project and customizing the tool for our need.	We still respond to questions and feature requests. Most likely not be able to dedicate enough time for any major coding efforts.	Y/N (willing to answer questions but not solving issues or implementing new features)	Y (eager to implement new features and improve existing ones)	Y (eager to implement new features and improve existing ones)
Other comments		- documents cannot be too big (need to split them)	- can be integrated with UCTO (tokenizer) to tokenize sentence or convert pre-tokenized sentence in FoLiA format - All changes history saved, user can revert them - document based (although you can activate sentence perspective where each sentence is separated from the next) - documents cannot be too big (need to split them)	- the interface is rather modest (basically a messaging app with only text and buttons) and cannot be made fancier, but depending how you look at it, it may be an advantage as it imposes extra pressure to make the logic of the system tidy and as simple as possible.

TOOL SELECTION PROCESS

FLAT

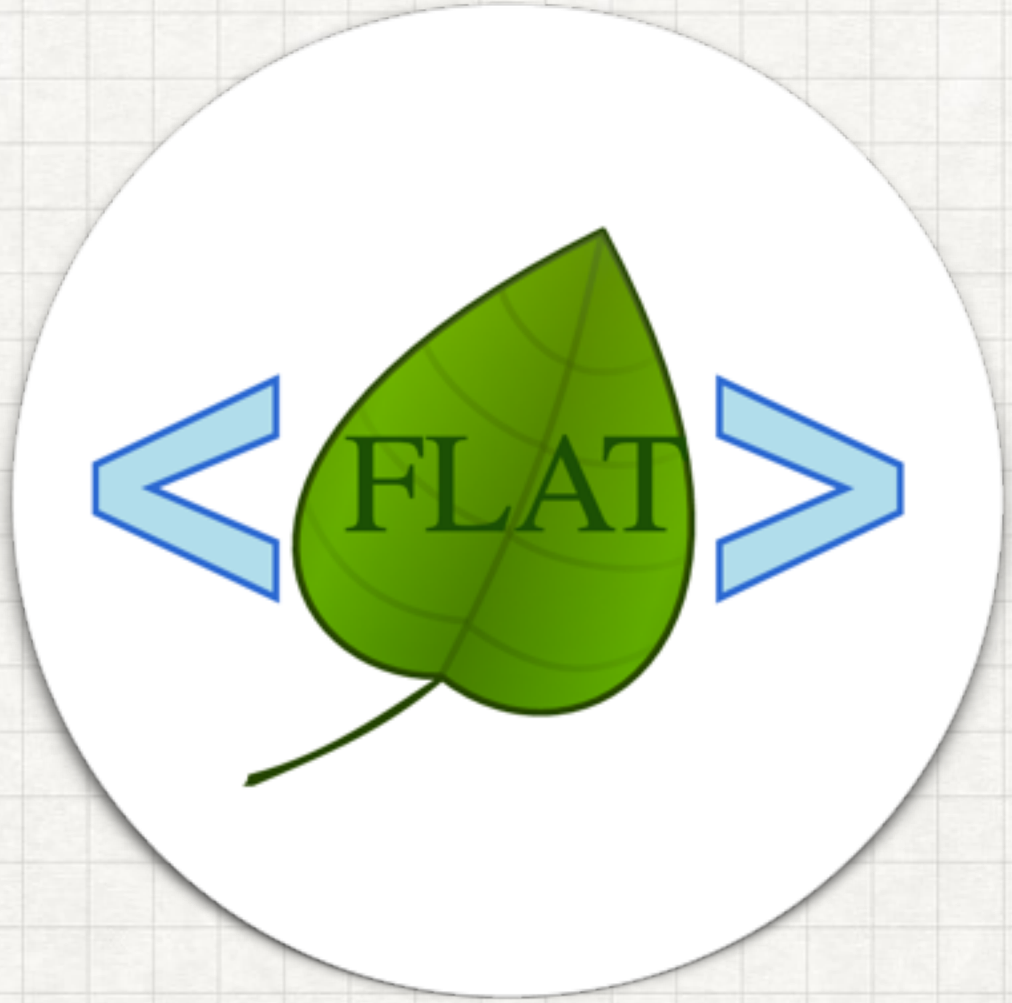
TELEGRAM-BOT



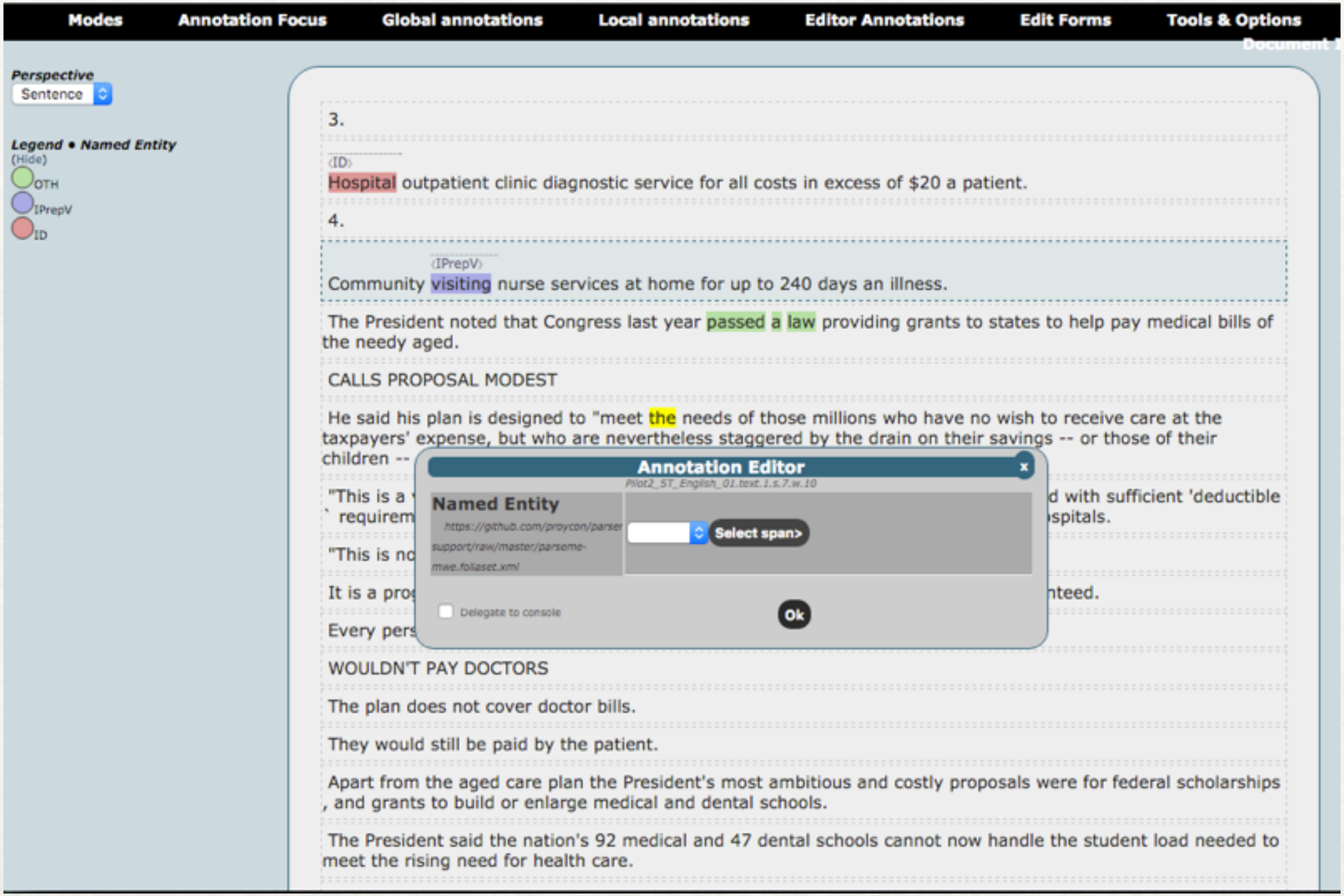


FLAT

- **FLAT**: web-application that offers an interface for the visualization and editing of FoLiA documents.
- **FoLiA**, a Format for Linguistic Annotation developed in the scope of the CLARIN-NL project and other projects.
- **Maarten van Gompel** (Centre for Language and Speech Technology Radboud Universiteit Nijmegen)



FLAT



3.

(ID)
Hospital outpatient clinic diagnostic service for all costs in excess of $20 a patient.

4.

(IPrepV)
Community visiting nurse services at home for up to 240 days an illness.

The President noted that Congress last year passed a law providing grants to states to help pay medical bills of the needy aged.

CALLS PROPOSAL MODEST

He said his plan is designed to "meet the needs of those millions who have no wish to receive care at the taxpayers' expense, but who are nevertheless staggered by the drain on their savings -- or those of their children --

"This is a … d with sufficient 'deductible ` requirem… spitals.

"This is no…

It is a pro… nteed.

Every pers…

Pilot2_ST_English_01.text.1.s.7.w.10

Named Entity

https://github.com/proycon/parser support/raw/master/parseme-mwe.foliaset.xml

□ Delegate to console

WOULDN'T PAY DOCTORS

The plan does not cover doctor bills.

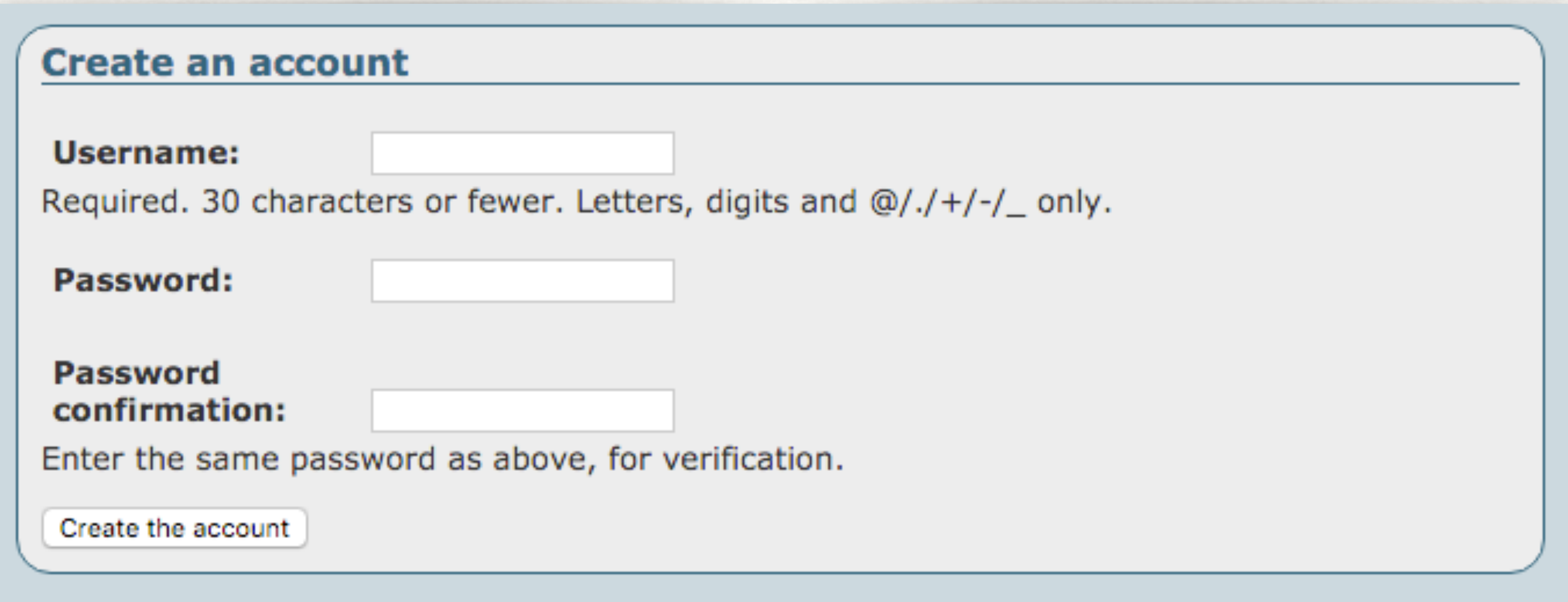
They would still be paid by the patient.

Apart from the aged care plan the President's most ambitious and costly proposals were for federal scholarships , and grants to build or enlarge medical and dental schools.

The President said the nation's 92 medical and 47 dental schools cannot now handle the student load needed to meet the rising need for health care.

FLAT

http://flat.science.ru.nl/register/



Create an account

Username:

Required. 30 characters or fewer. Letters, digits and @/./+/-/_ only.

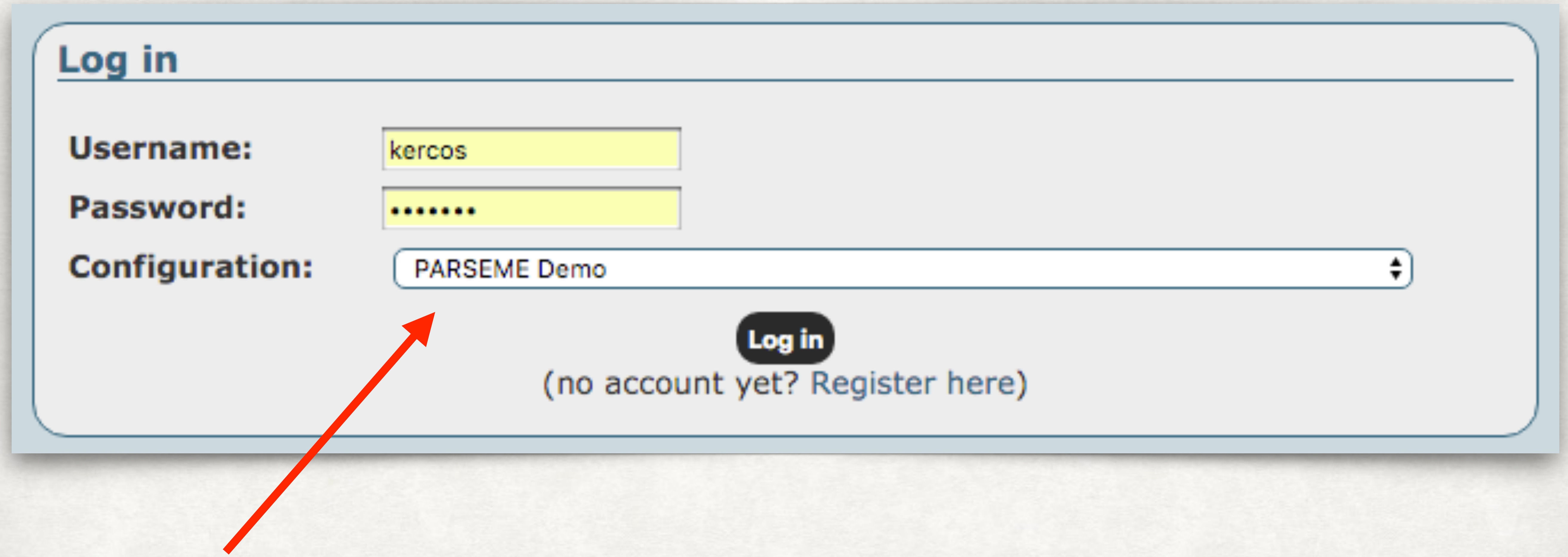
Password:

Password confirmation:

Enter the same password as above, for verification.

FLAT

http://flat.science.ru.nl/login



Log in

Username: kercos

Password: •••••••

Configuration: PARSEME Demo

FLAT

http://tiny.cc/flatCorpus

↓

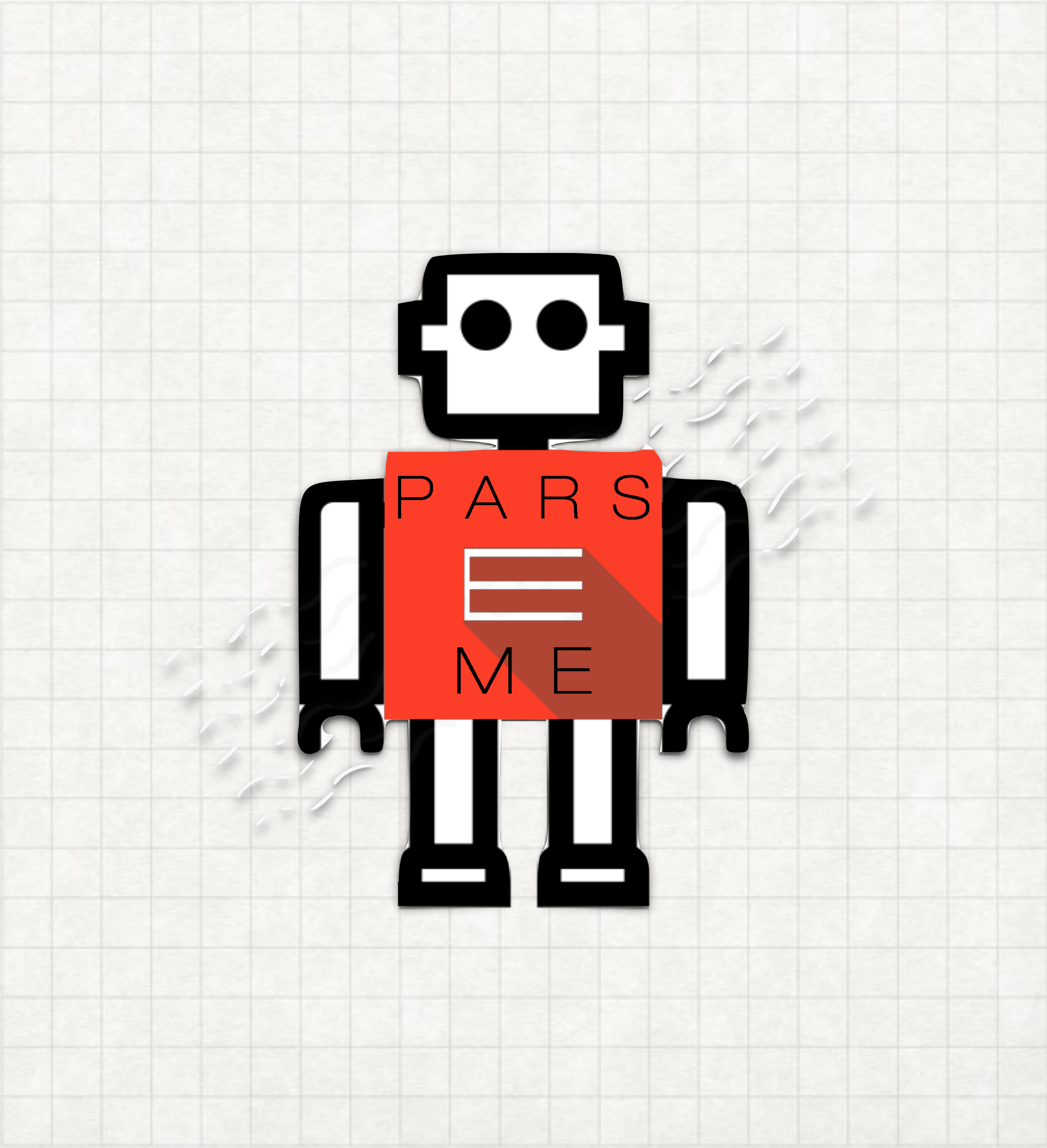
Pilot2_ST_English_FoLiA.zip

↓

Extract it into a local folder

TELEGRAM-BOT

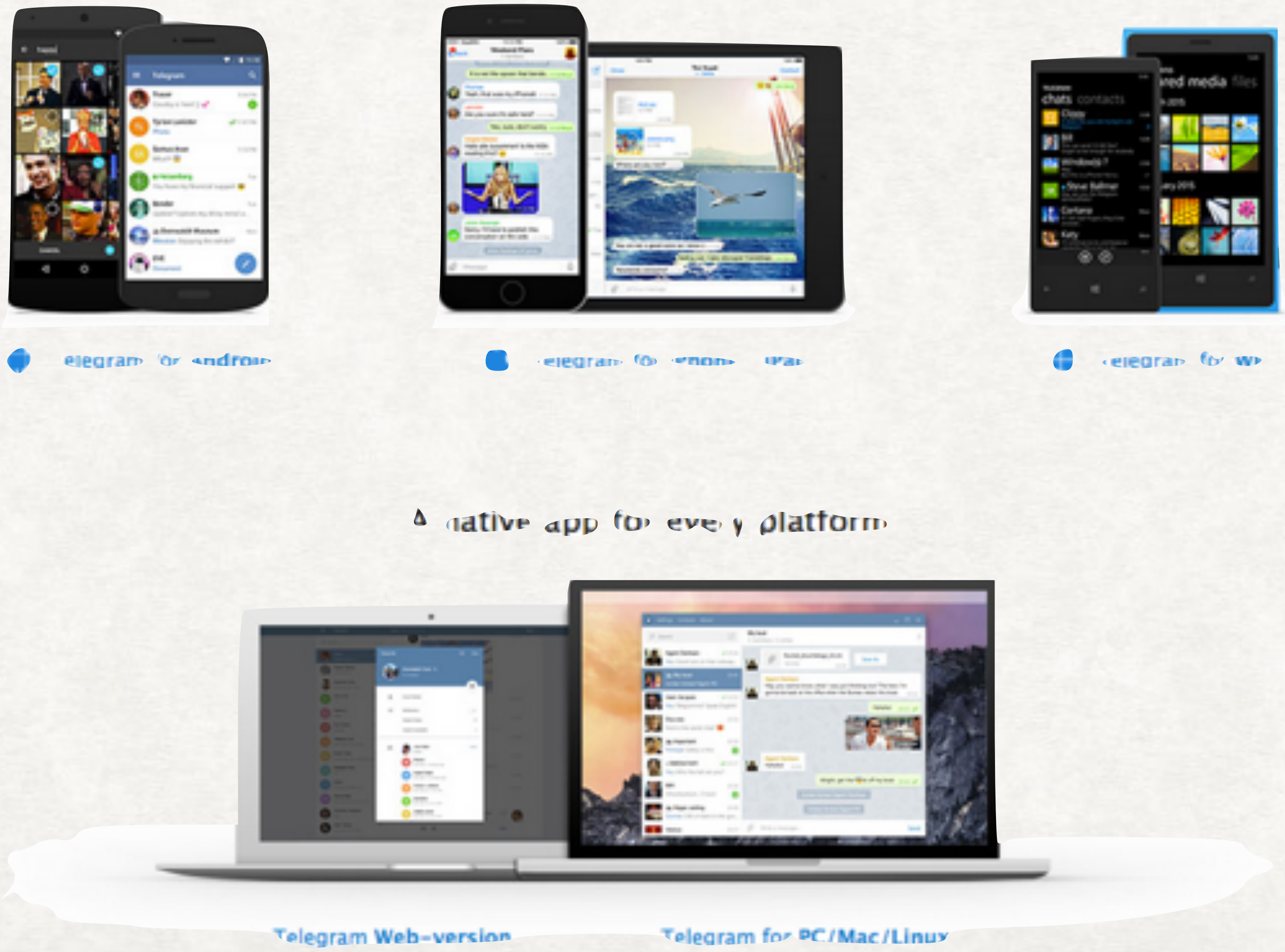
- In house application (Federico with the help of Behrang)
- Based on Telegram platform (messaging application)
- Text-based (simple interface with text and buttons)



TELEGRAM-BOT

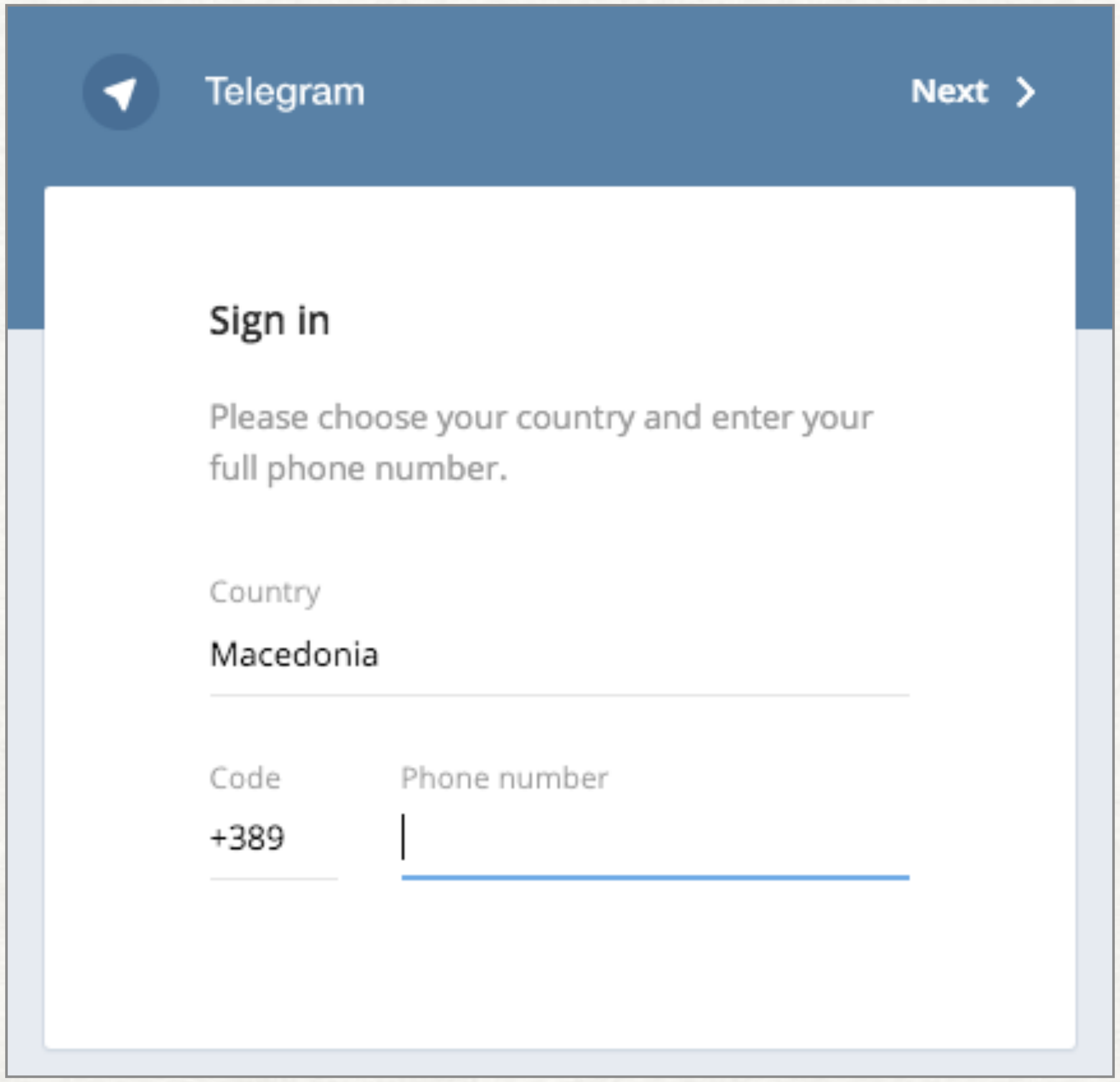
web.telegram.org

or install **Telegram app** in your PC/tablet/phone


Telegram for Android
Telegram for iPhone / iPad
Telegram for WP
A native app for every platform
Telegram Web-version
Telegram for PC/Mac/Linux

TELEGRAM-BOT

web.telegram.org



Telegram

Next >

Sign in

Please choose your country and enter your full phone number.

Country

Macedonia

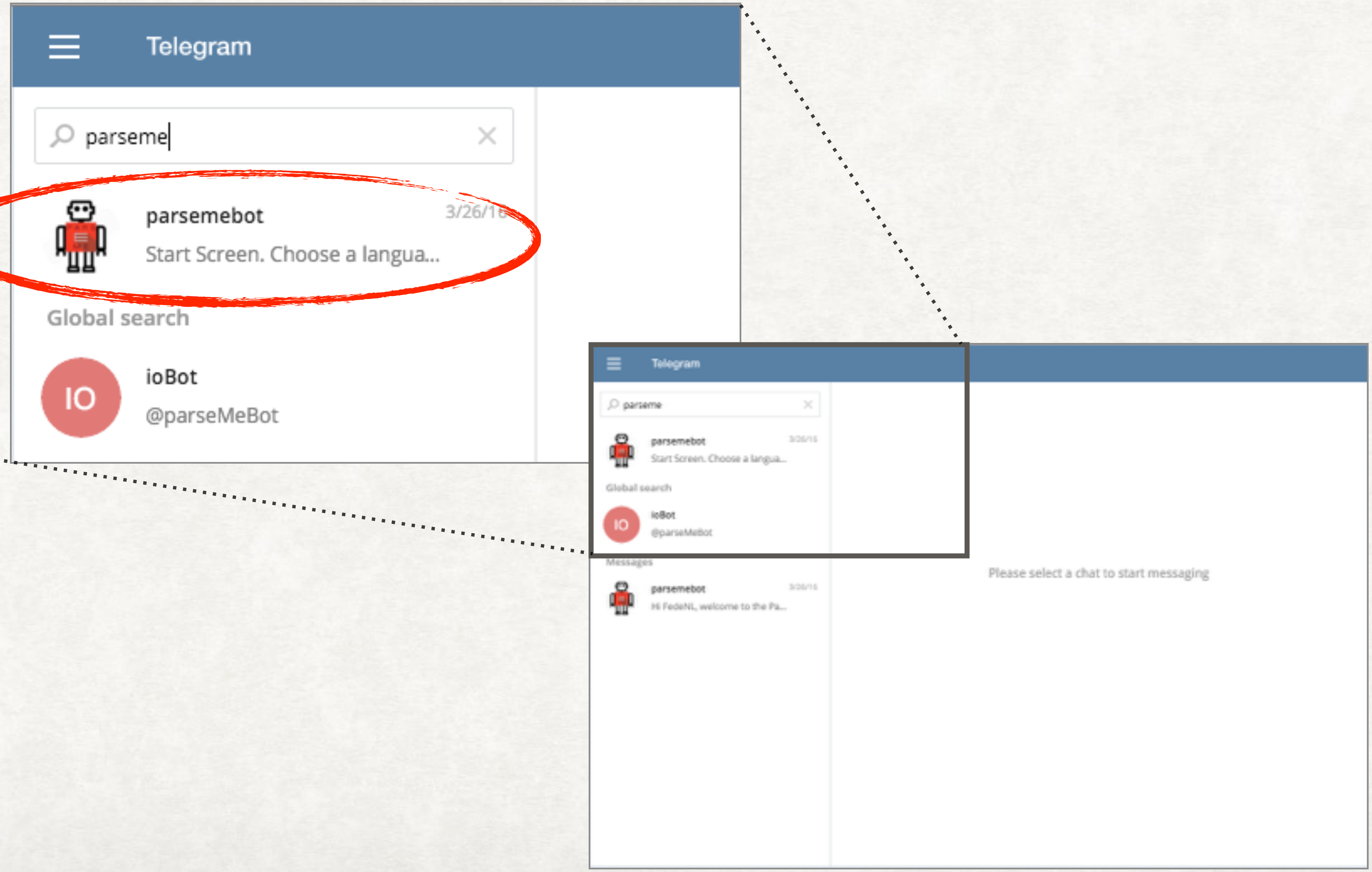
Code

+389

Phone number

TELEGRAM-BOT

web.telegram.org



Telegram

parseme

parsemebot 3/26/16
Start Screen. Choose a langua...

Global search

ioBot
@parseMeBot

Messages

parsemebot 3/26/16
Hi FedeNL, welcome to the Pa...

Please select a chat to start messaging

COMPARISON

	FLAT	TELEGRAM-BOT
CUSTOMIZATION	FEATURES COMMON TO OTHER ANNOTATION TASKS (E.G., RIGHT TO LEFT LANGUAGES)	ANY FEATURE SPECIFIC TO PARSEME TASK (E.G., ROLES, TRAINING PHASE, VALIDATION)
DEVICE COMPATIBILITY	WEB INTERFACE: ACCESSIBLE VIA A BROWSER	VIA TELEGRAM APP: ACCESIBLE VIA A BROWSER OR DEDICATED APP (PC, TABLET or PHONE
INTERFACE	DOCUMENT BASED (MULTIPLE SENTENCES)	SENTENCE BASED