Funded by the Horizon 2020 Framework Programme of the European Union
(COST Action 16105)

EnetCollect, a new COST Action of interest for the NLP community

(COST Action CA16105)

Games4NLP Workshop, LREC 2018

07/05/2018, Miyazaki, Japan

Verena Lyding & Lionel Nicolas

OVERVIEW

- Introduction
- Challenges, objectives & timeliness
- Network organization, composition & workplan
- Examples of foreseen implicit crowdsourcing for NLP
- Similarities & differences with GWAP approach
- Achievements up to now
- Next steps for the 2nd year

OVERVIEW

- **Introduction**
- Challenges, objectives & timeliness
- Network organization, composition & workplan
- Examples of foreseen implicit crowdsourcing for NLP
- Similarities & differences with GWAP approach
- Achievements up to now
- Next steps for the 2nd year

WHAT IS A COST ACTION ?

(1) Flexible networking instrument to cooperate, coordinate and jointly develop research ideas and funded research activities.

(2) Bottom-up driven networks supporting high-risk, innovative and emerging research themes.

(3) Active through workshops, conferences, training schools, short-term scientific missions (STSMs), and other dissemination activities.

(4) Up to ~ 900 000 euros over 4 years for network initiatives.
=> Does not fund research itself but paves the way to funded research.

⇒ COST Actions are a powerful networking means
to set into motion a new R&I trend

WHAT IS ENETCOLLECT ?

(1) "European NETwork for COmbining Language LEarning with Crowdsourcing Techniques".

(2) Performs the groundwork for a new R&I trend.

(3) Aims at unlocking a potential available for all languages by crowdsourcing on language learning and teaching activities to mass produce:
- language learning material (e.g. lessons or exercise content)
- language-related datasets (e.g NLP resources)

(4) Started in March 2017, will end in April 2021 (~4 years).

(5) International network involving ~185 stakeholders from Europe and outside.

EnetCollect welcomes new (proactive) members.

OVERVIEW

- ✓ **Introduction**
- ➢ **Challenges, objectives & timeliness**
- ➢ Network organization, composition & workplan
- ➢ Examples of foreseen implicit crowdsourcing for NLP
- ➢ Similarities & differences with GWAP approach
- ➢ Achievements up to now
- ➢ Next steps for the 2nd year

ENETCOLLECT'S CHALLENGES, OBJECTIVES & TIMELINESS

[CHALLENGES]

Long-term challenge

"Fostering the language skills of all citizens regardless of their backgrounds (social, linguistic, etc.) by enhancing the production of language learning material."

Short- to mid-term challenge

Incubating a new R&I trend to a point where multiple parallel and complementary finely-prepared projects relying on successful cooperations can be started.

=> EnetCollect does not aim at solving the problem but
aims at creating the R&I community that will solve the problem

ENETCOLLECT'S CHALLENGES, OBJECTIVES & TIMELINESS

[OBJECTIVES]

Research coordination - building shared knowledge

1. Creating a theoretical framework
2. Producing prototypical data
3. Disseminating the knowledge

Capacity building - creating a new and viable community

1. Forming a core community of stakeholders
2. Establishing communication channels
3. Obtaining new funded initiatives
4. Creating a stable association

ENETCOLLECT'S CHALLENGES, OBJECTIVES & TIMELINESS

[TIMELINESS]

EnetCollect is timely because of

(1) Increasing LL demand because of intensifying migration flows (e.g. market globalization, political developments, etc.)
- Always more language learners
- More diversified target groups but LL material produced is at country-level
- Combinations languages / target groups requires a large-scale approach

(2) Favorable conditions for combining crowdsourcing and LL on a large scale
- Crowdsourcing is now omnipresent in language-related R&I fields
- Language-related R&I fields make very limited use of LL to crowdsource
- Funding agencies have acknowledged crowdsourcing

OVERVIEW

- ✓ **Introduction**
- ✓ **Challenges, objectives & timeliness**
- ➢ **Network organization, composition & workplan**
- ➢ Examples of foreseen implicit crowdsourcing for NLP
- ➢ Similarities & differences with GWAP approach
- ➢ Achievements up to now
- ➢ Next steps for the 2nd year

NETWORK ORGANIZATION, COMPOSITION & WORKPLAN

[ORGANIZATION]

5 Working Groups	Primary WG / Mailing List
• WG1: Explicit Crowdsourcing for LL material production	41 / 62
• WG2: Implicit Crowdsourcing for LL material production	31 / 61
• WG3: User-oriented design strategies for a competitive solution	23 / 51
• WG4: Technology-oriented specs for a flexible/robust solution	12 / 33
• WG5: Application-orient specs for an ethical, legal and profitable solution	5 / 30

3 transversal groups	Mailing list
• Outreach Coordination	10
• Exploitation Coordination	14
• Dissemination coordination	16

NETWORK ORGANIZATION, COMPOSITION & WORKPLAN

[COMPOSITION]

Members with heterogeneous background

(1) Crowdsourcing

(2) Computer Assisted Language Learning

(3) Natural Language Processing

(4) E-lexicography

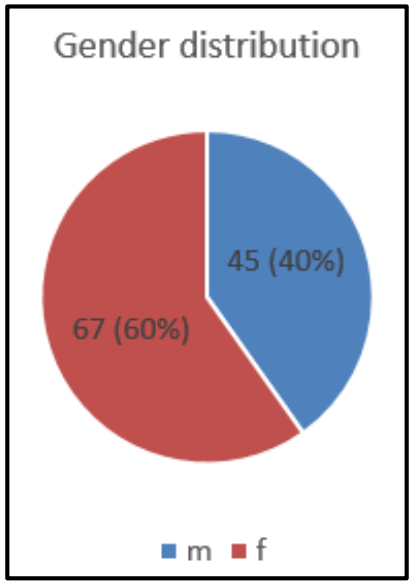
(5) Learning Management Systems

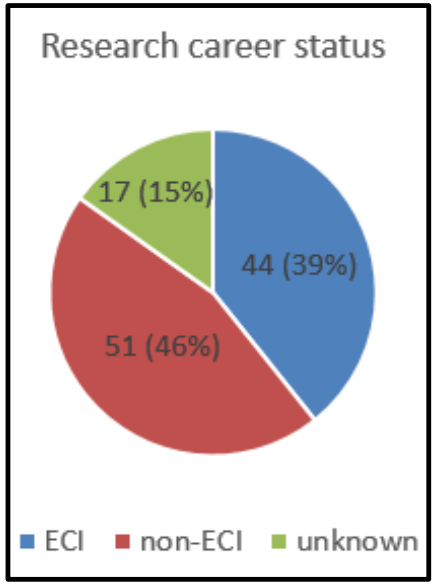
(6) Learner Corpora

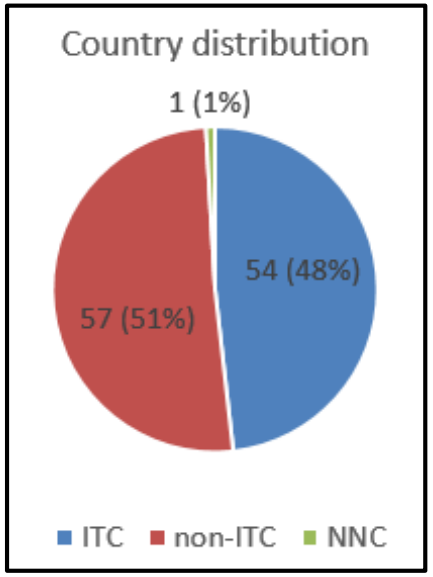
(7) Corpus Linguistics

NETWORK ORGANIZATION, COMPOSITION & WORKPLAN

[COMPOSITION]







NETWORK ORGANIZATION, COMPOSITION & WORKPLAN

[WORKPLAN]

WGs / Months	6	12	18	24	30	36	42	48
WG 1 + 2 + 3	State-of-the-art							
		Brainstorming						
				Prototyping				
WG 4 + 5		Guidelines, technical solutions, blueprints						
OED	IPR and OED plans, communication means and dissemination							

OVERVIEW

- ✓ **Introduction**
- ✓ **Challenges, objectives & timeliness**
- ✓ **Network organization, composition & workplan**
- ➢ **Examples of foreseen implicit crowdsourcing for NLP**
- ➢ Similarities & differences with GWAP approach
- ➢ Achievements up to now
- ➢ Next steps for the 2nd year

Examples of foreseen implicit crowdsourcing for NLP

[CORE CONCEPTS]

(1) Implicit crowdsourcing with LL exercises is most adequate for crowdsourcing NLP

(2) LL exercises used need to be meaningful in order to
- not waste the learners' efforts to learn
- not drive away the learners from the language learning platform

(3) Overall logic is => if NLP resources can be used to generate exercise content
Then learners' answers can be used to correct or extend the NLP resource

(4) Every crowdsourced answer is a win-win
=> the NLP researchers obtains new data
=> the learners (as a whole) obtain more exercise content

Examples of foreseen NLP implicit crowdsourcing

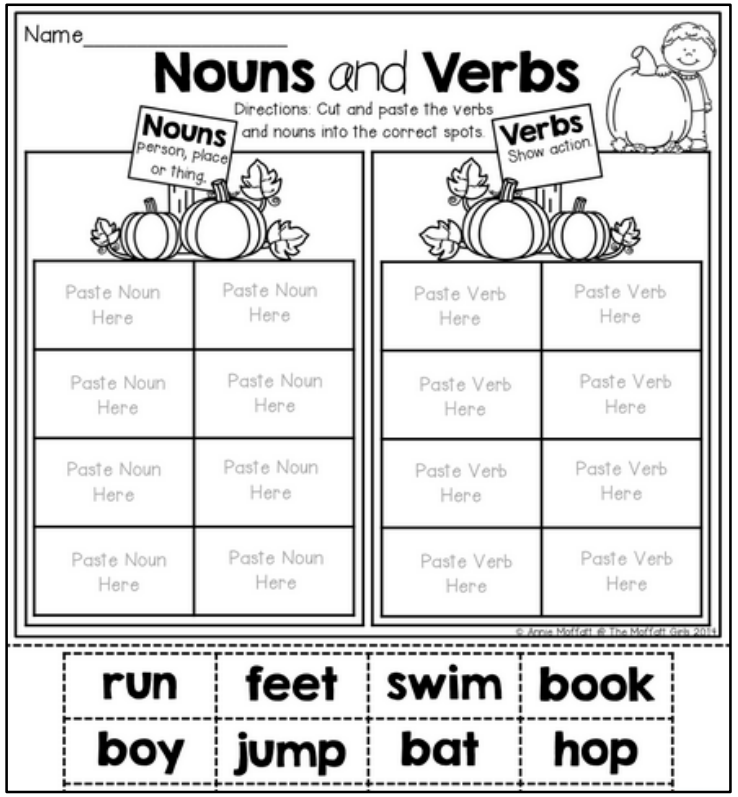
[CROWDSOURCING OF LEXICAL DATA]

"Word search" exercises

S	E	I	R	O	L	A	C	I	T	P	I	L	L	E
T	R	A	I	N	E	R	S	T	H	G	I	E	W	R
N	O	I	S	N	E	T	E	N	N	I	S	E	A	U
I	S	W	A	T	E	R	D	O	P	I	S	L	L	S
R	C	B	S	T	R	E	T	C	H	W	U	C	K	S
P	I	U	L	A	S	H	I	K	E	C	F	S	C	E
S	N	L	L	R	T	I	Y	A	S	L	I	U	I	R
U	E	C	I	G	A	M	T	A	L	A	T	M	B	P
O	H	I	M	E	M	B	V	C	U	O	W	I	O	D
U	T	B	D	T	G	O	J	W	P	G	O	K	R	O
N	S	O	A	M	I	W	S	Y	N	U	R	S	E	O
E	I	R	E	D	E	N	E	R	G	Y	M	C	A	L
R	L	E	R	C	O	O	L	D	O	W	N	R	N	B
T	A	A	T	R	O	U	T	I	N	E	C	N	A	D
S	C	H	E	D	U	L	E	T	U	O	K	R	O	W

Copyright 2007 John R. Potter John's Word Search Puzzles
http://www.thepotters.com/puzzles.html

"Classify words" exercises

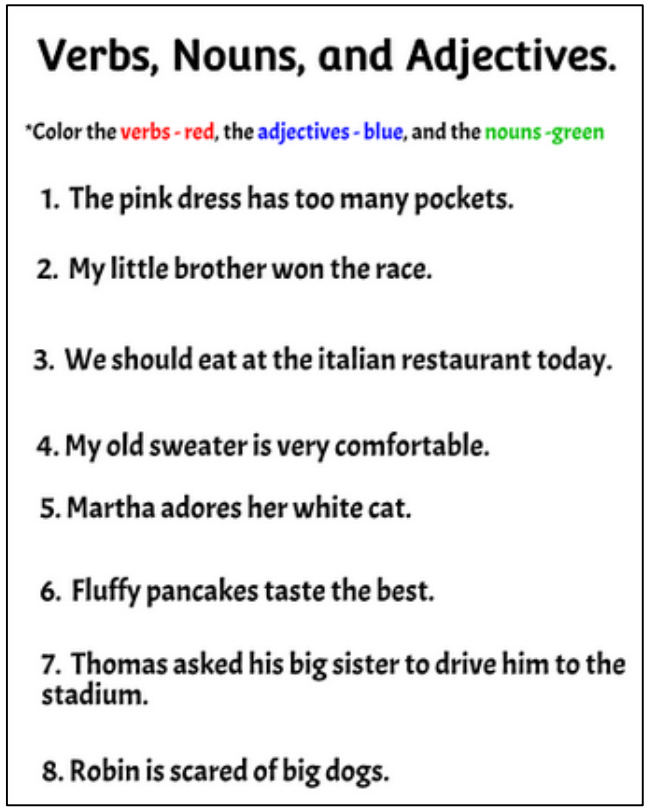


© https://www.pinterest.com/pin/153052087312404564/

Examples of foreseen NLP implicit crowdsourcing

[CROWDSOURCING OF GRAMMATICAL DATA]

"Color the word" exercises

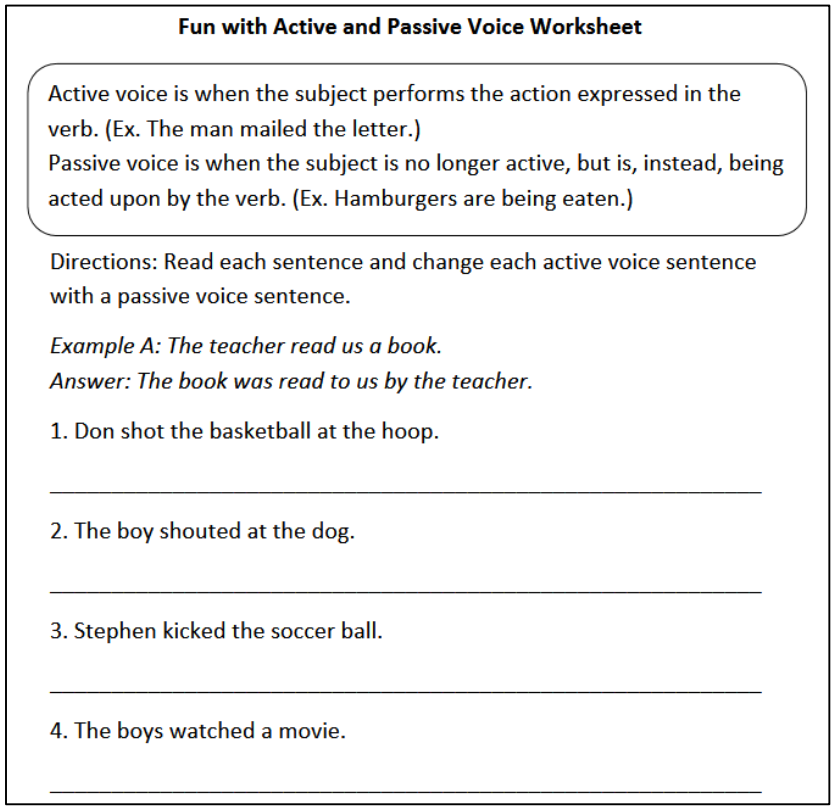


Verbs, Nouns, and Adjectives.

*Color the verbs - red, the adjectives - blue, and the nouns -green

1. The pink dress has too many pockets.
2. My little brother won the race.
3. We should eat at the italian restaurant today.
4. My old sweater is very comfortable.
5. Martha adores her white cat.
6. Fluffy pancakes taste the best.
7. Thomas asked his big sister to drive him to the stadium.
8. Robin is scared of big dogs.

"Passive / Active voice" exercise



Fun with Active and Passive Voice Worksheet

Active voice is when the subject performs the action expressed in the verb. (Ex. The man mailed the letter.)
Passive voice is when the subject is no longer active, but is, instead, being acted upon by the verb. (Ex. Hamburgers are being eaten.)

Directions: Read each sentence and change each active voice sentence with a passive voice sentence.

Example A: The teacher read us a book.
Answer: The book was read to us by the teacher.

1. Don shot the basketball at the hoop.

_____

2. The boy shouted at the dog.

_____

3. Stephen kicked the soccer ball.

_____

4. The boys watched a movie.

_____

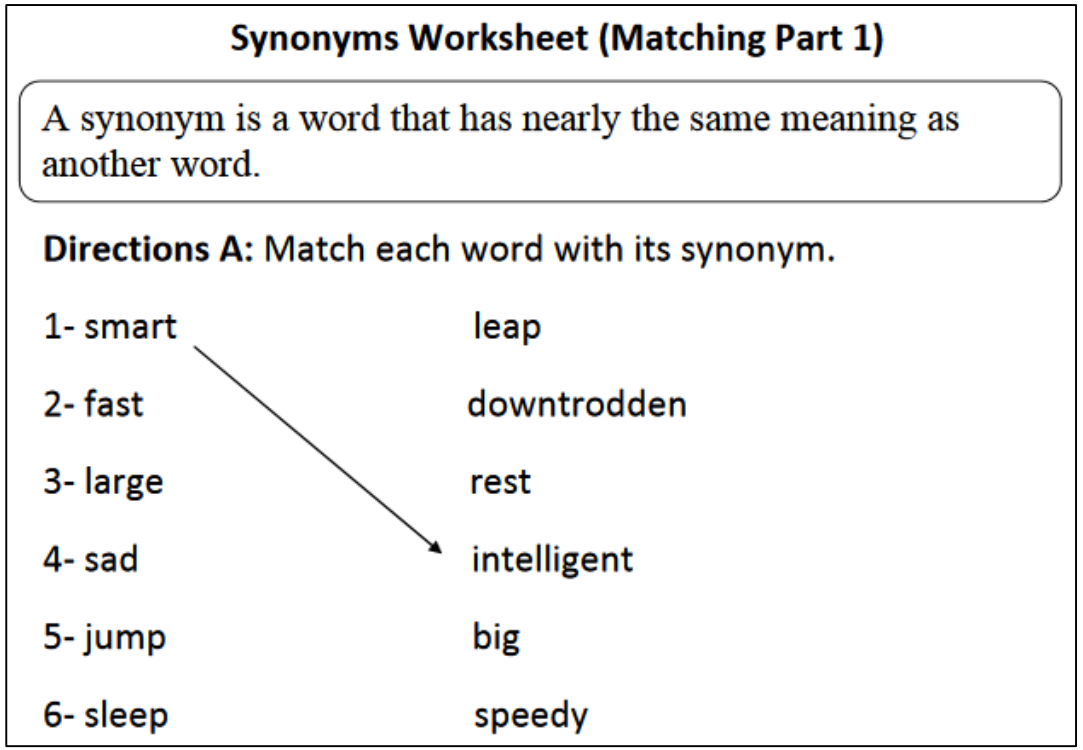
Examples of foreseen NLP implicit crowdsourcing

[CROWDSOURCING OF SEMANTIC DATA]

"Analogy" exercises

1. Happy is to Joyful as Sad is to _____.
2. Loud is to Noisy as Quiet is to _____.
3. Yellow is to Corn as Green is to _____.
4. Pen is to Writer as Voice is to _____.
5. Fly is to Airplane as Drive is to _____.
6. Artist is to Painting as Baker is to _____.
7. Beagle is to Dog as Canary is to _____.
8. Scissor is to Cut as Ruler is to _____.
9. Wheel is to Circle as Book is to _____.
10. Hat is to Head as Sneaker is to _____.

"Synonymy" exercises



Synonyms Worksheet (Matching Part 1)

A synonym is a word that has nearly the same meaning as another word.

Directions A: Match each word with its synonym.

1- smart	leap
2- fast	downtrodden
3- large	rest
4- sad	intelligent
5- jump	big
6- sleep	speedy

Examples of foreseen NLP implicit crowdsourcing

[POTENTIAL EVALUATION]

Hypothesis

(1) We only consider European learners over 14 (~ 80 millions)

(2) We have developed “Wordpress-like” platforms that researchers can implement

(3) Crowdsourcing 1 learner over 1 year = 1 hour of expert manpower

(4) An expert works ~1900 hours per year (5 days out of 7 days, 47 weeks out of 52)

(5) The average cost of expert manpower ~35 000 euros per year (~ 18,5 euros per hour)

- Crowdsourcing 1% of 80 million learners => ~ 15 000 000 euros
- Crowdsourcing 10% of 80 million learners => ~ 150 000 000 euros
- Crowdsourcing 100% of 80 million learners => ~ 1 500 000 000 euros

OVERVIEW

- ✓ **Introduction**
- ✓ **Challenges, objectives & timeliness**
- ✓ **Network organization, composition & workplan**
- ✓ **Examples of foreseen implicit crowdsourcing for NLP**
- ➢ **Similarities & differences with GWAP approach**
- ➢ Achievements up to now
- ➢ Next steps for the 2nd year

SIMILARITIES & DIFFERENCES WITH THE GWAP APPROACH

[SIMILARITIES]

Both GWAP and COLLECT approaches for NLP resource creation

(1) Collect data primarily through implicit crowdsourcing

(2) Tend to split tasks into microtasks

(3) Need to evaluate the crowd's reliability over time

(4) Have to ensure to attract and retain the crowd

(5) Have a potentially infinite crowd that can get always self-renewed

SIMILARITIES & DIFFERENCES WITH THE GWAP APPROACH

[DIFFERENCES]

GWAP targets a crowd of players
COLLECT targets a crowd of language learners & teachers
⇒ Different scale of crowdsourcing potential

GWAP is online only
COLLECT is online but can also rely on local (political) support
⇒ Local support can make a noticeable difference

GWAP requires to devise games that are fun
COLLECT requires to devise exercises that are relevant (fun is a +)
⇒ The crowd's expectations can be easier to address for COLLECT

SIMILARITIES & DIFFERENCES WITH THE GWAP APPROACH

[DIFFERENCES]

GWAP requires to find a balance between testing and crowdsourcing
COLLECT crowdsources rarely and tests (i.e. teaches) most of the time
=> Reliability of the crowd is easier to evaluate for COLLECT
=> Crowdsourcing is more profitable for GWAP in terms of user time

GWAP targets the "free" time of the crowd (i.e. shorter but intense timespans)
COLLECT targets the "study" time of the crowd (i.e. longer but "calm"timespans)
⇒ The number of hours per person in the crowd should be larger for COLLECT

GWAP competes with numerous innovative solutions
COLLECT competes with not numerous, hardly innovative but well established solutions
⇒ The crowd's size is more volatile for GWAP
⇒ The crowd's size is less volatile for COLLECT (but initial efforts are especially demanding)

OVERVIEW

- ✓ **Introduction**
- ✓ **Challenges, objectives & timeliness**
- ✓ **Network organization, composition & workplan**
- ✓ **Examples of foreseen implicit crowdsourcing for NLP**
- ✓ **Similarities & differences with GWAP approach**
- ➢ **Achievements up to now**
- ➢ Next steps for the 2nd year

ACHIEVEMENTS UP TO NOW

[NUMBER OF MEMBERS, COUNTRIES & BUDGET]

	03/17	09/17	03/18
Members' presence	~100 on Mailing list (no intranet)	~150 on Mailing list ~85 on Intranet	~185 on Mailing list ~115 on Intranet
Representatives	~ 70	99	109
Countries	28	34	38
Budget	173 000	188 000	210 000

ACHIEVEMENTS UP TO NOW

[MEETINGS & TYPE OF MEETINGS]

9 meetings organized of 5 different types

- 3 Action Meetings
  - ✓ Kick-off meeting in Brussels (03/17, 1 day, 47 persons)
  - ✓ 1st Annual meeting in Bolzano (09/17, 2 days, 55 persons)
  - ✓ 2nd Annual meeting in Iasi (03/18, 3 days, 75 persons)
- 1 Training School in Iasi (03/08)
- 1 Core Group meeting in Ljubljana (06/17)
- 4 online meetings

ACHIEVEMENTS UP TO NOW

[STSMS, SCIENTIFIC PUBLICATIONS & DISSEMINATION]

- 12 STSMs with 14 members as STSM grantees or hosts
- 3 submissions describing the overall ambition of the enetCollect Action
- 2 invited presentations
- Numerous press releases
- Website & intranet
- 11 mailing lists

ACHIEVEMENTS UP TO NOW

[PROJECT FUNDING]

Initiatives to foster project proposal writing

- List of “autonomous” funding opportunities created.
- Initiative to foster the sharing of project proposals.
- One campaign to foster MCIF proposals.

COST-related funding schemes

- Partners at Uni. Geneva got a mid-sized 3-years project (explicit crowdsourcing).

MCIF

- Partner at EURAC (us) got a rejected application with a seal of excellence (implicit crowdsourcing). Potential funding by local authorities (to be confirmed).

OVERVIEW

- ✓ Introduction
- ✓ Challenges, objectives & timeliness
- ✓ Network organization, composition & workplan
- ✓ Examples of foreseen implicit crowdsourcing for NLP
- ✓ Similarities & differences with GWAP approach
- ✓ Achievements up to now
- ➢ **Next steps for the 2nd year**

NEXT STEPS FOR THE 2ND YEAR

Work orientations / Phases

Some state of the art, a lot of brainstorming and a bit of prototyping.

[COST tools]

- 15 ~ 20 STSMs
- 5 WG Meetings late 2018 (some collocated), 70 ~ 90 invitations in total
- 1 Annual Meeting early 2019 (in Lisbon), 80 ~100 invitations
- 1 or 2 Training Schools
- 1 Hackathon

[Dissemination, exploitation & varia]

- 1 coordinated campaign to foster MCIF proposal
- 1 coordinated campaign to foster Erasmus key Action 2 proposal
- 1 coordinated campaign to foster community-oriented publications
- Many many many many other things….

Many thanks for your attention!
Questions ?

Join! => http://enetcollect.eurac.edu/joining-enetcollect/

Contact: chair.enetcollect@eurac.edu