Slide 1

HEALTH CARE | UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

6TH WORLD CONGRESS ON FLUENCY DISORDERS
5th – 8th August, 2009 Rio de Janeiro, Brazil

ifa international fluency association

Timing and Tallying Dysfluencies

using *Praat* computer software

Paul Corthals[a,b]

[a] Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[b] Faculty of Health Care 'Vesalius', University College Ghent, Ghent, Belgium

UNIVERSITEIT GENT

I want to fuse acoustic phonetics and stuttering. The idea is that stuttering is a prosodic phenomenon and prosody can be measured using the tools of acoustic phonetics.

Slide 2

HEALTH CARE | UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

PART I
Dysfluency taxonomies
PART II
How to implement a taxonomy in Praat
PART III
Detailed case study and discussion

2

UNIVERSITEIT GENT

Slide 4

HEALTH CARE | UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

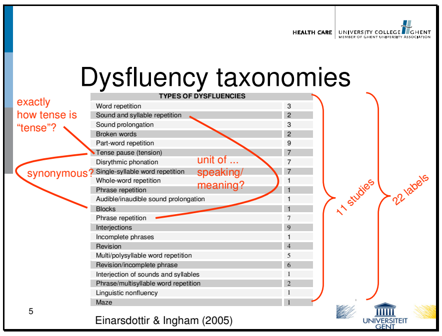
Clonic
Tonic
Primary
Secondary
Word repetition
Sound repetition
Syllable repetition
Interjection
Incomplete prase
Prolongation
Revision
Disryhtimic phonation
Part-word repetition

4

UNIVERSITEIT GENT

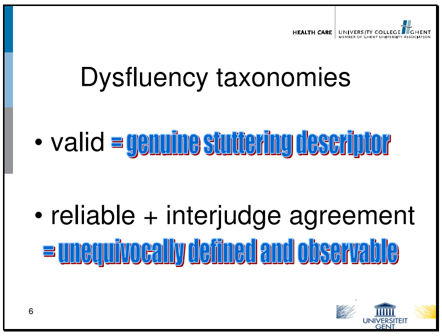
First let's define what we are studying. We are focusing on dysfluency types. A lot of jargon is used in stuttering literature.

Slide 5



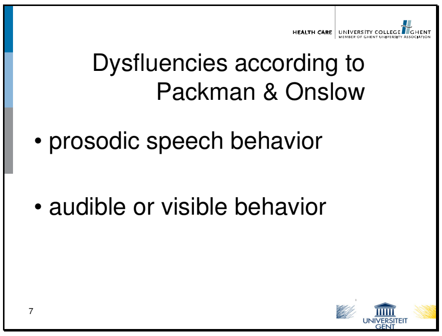
This list of dysfluencies comes from a review by Einarsdottir and Ignham. Clearly, there is a plethora of labels, which does not help students, nor reseachers for that matter. Some terms seem synonymous, but they may convey a different nuance each (tense pause & blocks, broken words & part-word repetition). Some refer to a unit of meaning or a unit of syntax, suggesting that semantics or syntactics are implicated in the disorder and confounding cross-linguistic evaluation (incomplete phrases, linguistic nonfluency, part-word repetition, broken words), in contrast to labels refering to syllables, which are motor speech units.

Slide 6



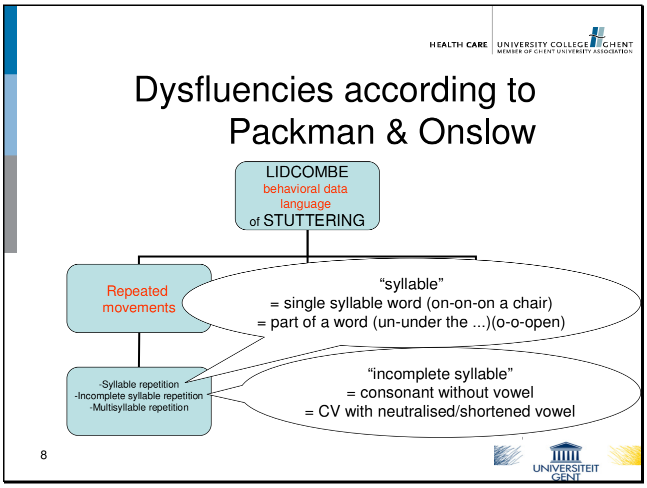
Dysfluency taxonomies are meant to test the type and severity of the stuttering or cluttering problem. The basic requirements for any testing tool, whatever it measures, are validity and reliablility. A test is valid when it actually measures what it promises to measure. In the case of a taxonomy of stuttering dysfluencies this means...

Slide 7

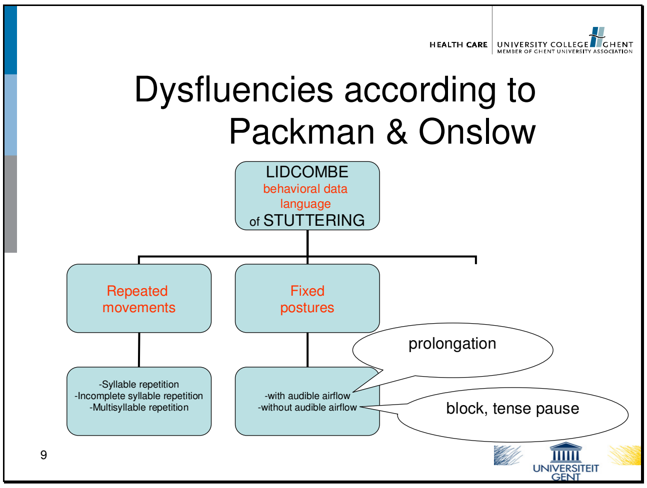


The best guarantee for validity and reliability is selecting descriptors that can be defined unequivocally. When we stick to prosodic speech features , i.e. observing articulation manoeuvres in time we can be more concrete. This is not to say there are no other stuttering symptoms beyond the realm of motor speech (rephrasing, avoiding certain words), but those are less observable, less measurable (I warned you, I am a phonetician addicted to acoustics). Packman and Onslow rearranged the most robust stuttering descriptors into a simple framework.

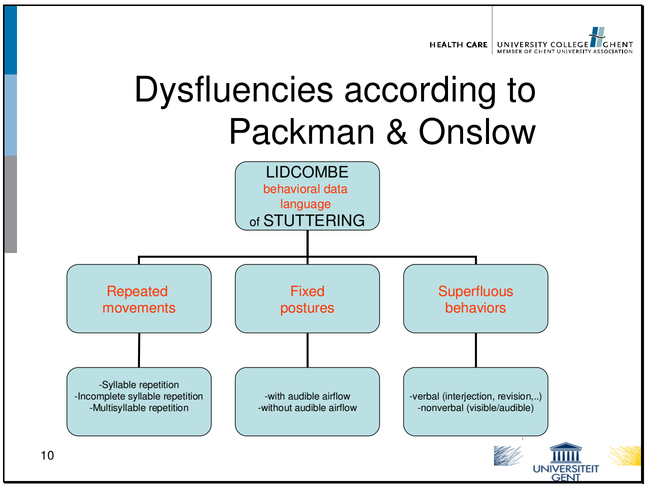
Slide 8



Slide 9



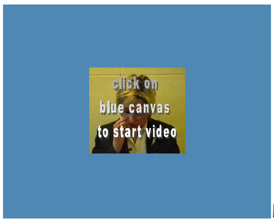
Slide 10



Perhaps this model requires considerable reconceptualization when you are a seasoned stuttering expert, but the descriptors seem comprehensive without being redundant. Let's do a try-out.

Slide 11

Dysfluencies according to Packman & Onslow



case #1

VIDEO FRAGMENT (click on the blue canvas)

Slide 12

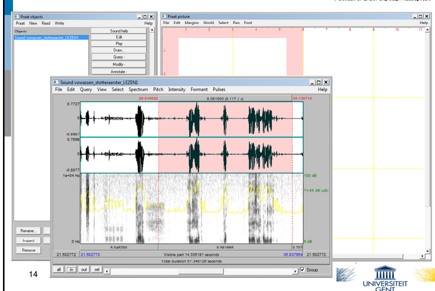
PART I

Dysfluency taxonomies

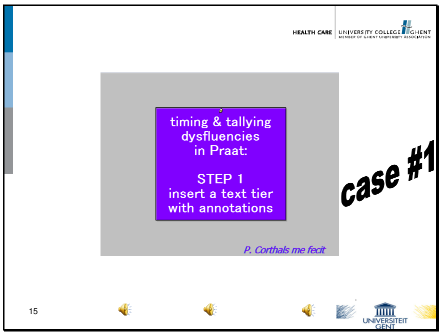
PART II

How to implement a taxonomy in Praat

Slide 14



Slide 15

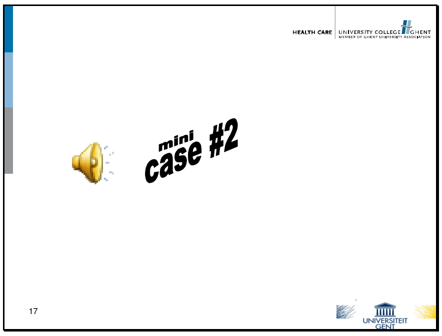


ANIMATION (click on the blue canvas)

Slide 16

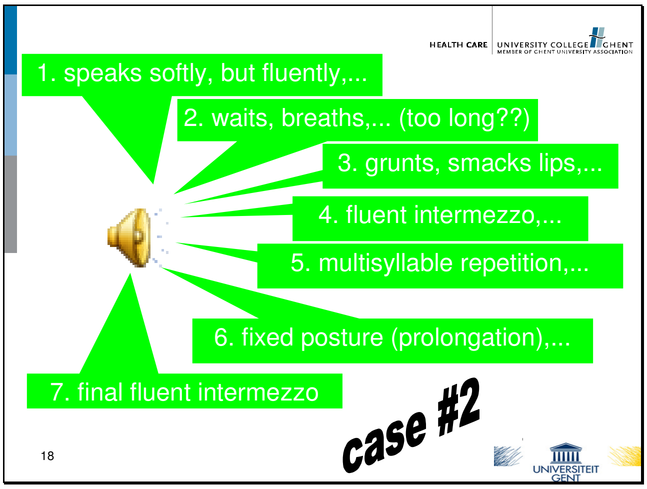
Superfluous behavior, if nonverbal and inaudible, does not appear on a Praat editor screen. If we want to time it exactly, we could add a separate TextGrid for inaudible nonverbal superfluous behavior.

Slide 17



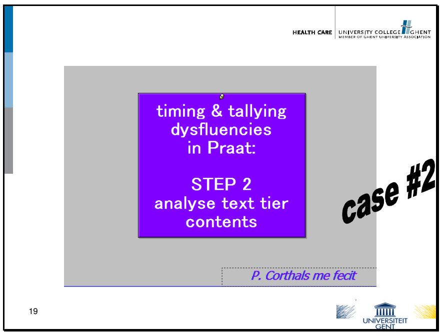
Listen carefully and try to recognize the dysfluency types in the sample. Relax: it is just a 10 second sample.

Slide 18



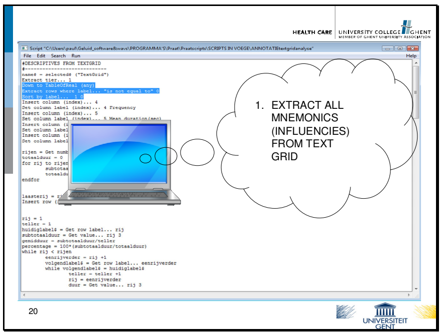
Those long inhalation pauses: are they stuttering symptoms?

Slide 19



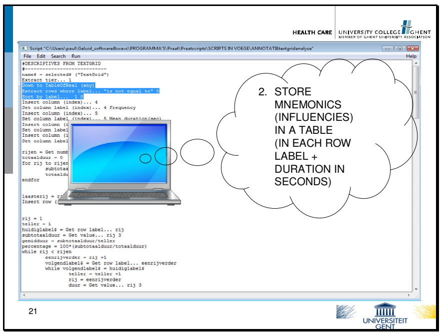
We have to annotate the text grid, and then how do we proceed? ANIMATION (click on the blue canvas)

Slide 20

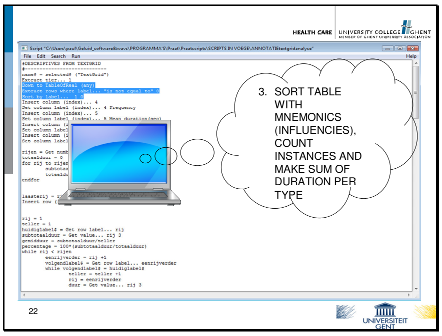


A Praat script isn't Voodoo. Once you have it, it works fast and in a reproducible way.

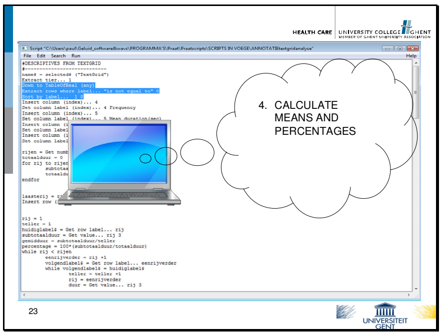
Slide
21



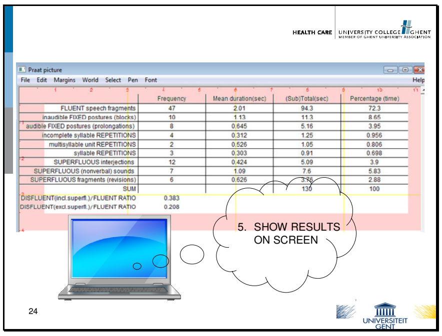
Slide
22



Slide
23

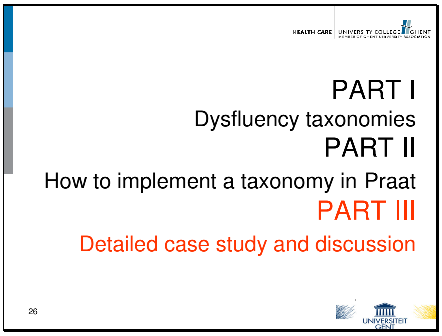


Slide 24

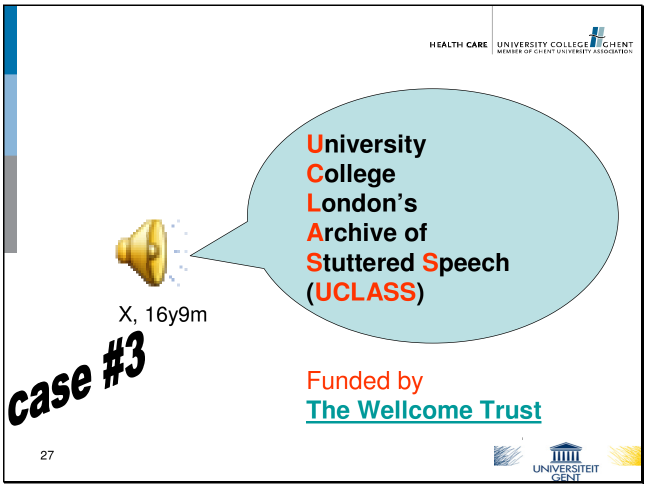


	Frequency	Mean duration(sec)	(Sub)Total(sec)	Percentage (time)
FLUENT speech fragments	47	2.01	94.3	72.3
inaudible FIXED postures (blocks)	10	1.13	11.3	8.65
audible FIXED postures (prolongations)	8	0.645	5.16	3.95
incomplete syllable REPETITIONS	4	0.312	1.25	0.956
multisyllable unit REPETITIONS	2	0.526	1.05	0.806
syllable REPETITIONS	3	0.303	0.91	0.698
SUPERFLUOUS interjections	12	0.424	5.09	3.9
SUPERFLUOUS (nonverbal) sounds	7	1.09	7.6	5.83
SUPERFLUOUS fragments (revisions)	6	0.626	3.76	2.88
SUM			130	100
DISFLUENT(incl.superfl.)/FLUENT RATIO	0.383			
DISFLUENT(excl.superfl.)/FLUENT RATIO	0.208			

Slide 26

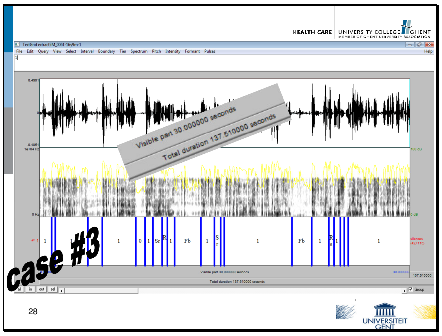


Slide 27



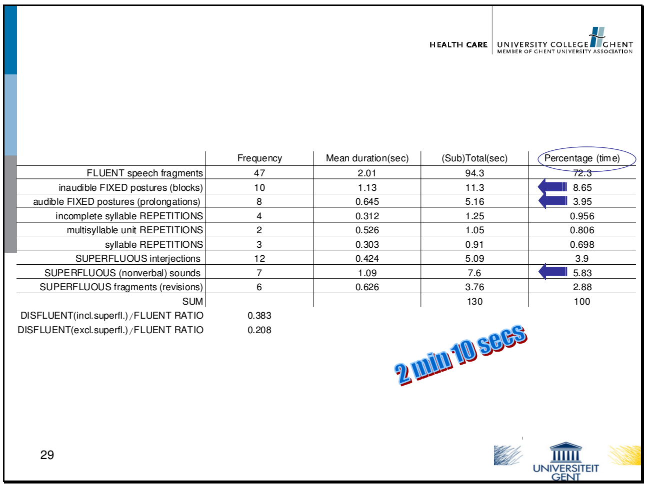
We will analyze case #3 in a detailed way. It is a two-minute sample that can be downloaded from the website of UCLASS. Listen to it.

Slide 28



We have annotated the sample in Praat. Of course, this was a personal view. However, Packman & Onslows' taxonomy is rather straightforward, so I believe another person would pretty much end up with the same categories. Nevertheless, I think the inter-rater reliability of annotations could improve if all users could follow the same training before they start.

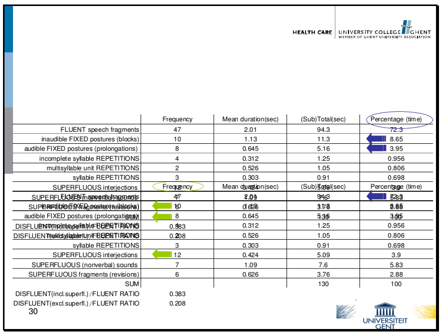
Slide 29



	Frequency	Mean duration(sec)	(Sub)Total(sec)	Percentage (time)
FLUENT speech fragments	47	2.01	94.3	72.3
inaudible FIXED postures (blocks)	10	1.13	11.3	8.65
audible FIXED postures (prolongations)	8	0.645	5.16	3.95
incomplete syllable REPETITIONS	4	0.312	1.25	0.956
multisyllable unit REPETITIONS	2	0.526	1.05	0.806
syllable REPETITIONS	3	0.303	0.91	0.698
SUPERFLUOUS interjections	12	0.424	5.09	3.9
SUPERFLUOUS (nonverbal) sounds	7	1.09	7.6	5.83
SUPERFLUOUS fragments (revisions)	6	0.626	3.76	2.88
SUM			130	100
DISFLUENT(incl.superfl.)/FLUENT RATIO	0.383			
DISFLUENT(excl.superfl.)/FLUENT RATIO	0.208			

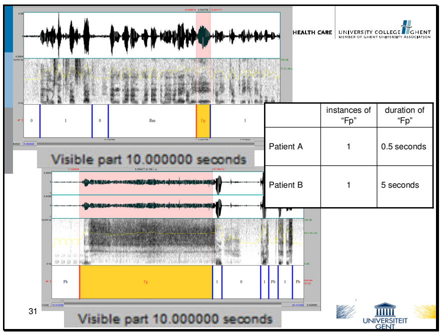
These are the results. The top three descriptors reveal the profile of a stutterer who predominantly has blockades, and prolongations (both fixed postures) and a lot of superflous nonverbal sounds and interjections.

Slide 30



Note that the interjections can be numerous (#1 in frequency count) but short (#3 in percentage of time). This raises the question what measure correlates best with a listeners' impression of stuttering severity: frequency counts of timing percentages?

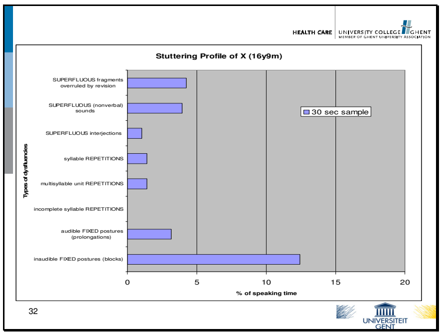
Slide 31



	instances of "Fp"	duration of "Fp"
Patient A	1	0.5 seconds
Patient B	1	5 seconds

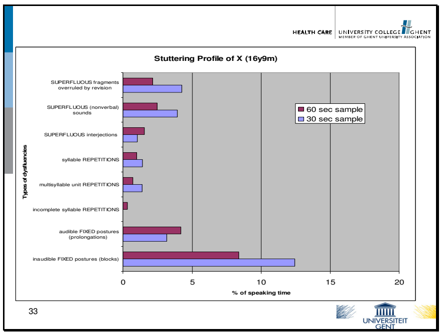
For fixed postures (blocks or prolongations) there is a crucial difference between tallying and timing. Here you see exactly one prolongation "Fp" on both screens, so patient A and patient B have the same frequency of prolongations. Both screens cover exactly 10 seconds of speech. So two patients can have the same frequency of fixed postures, but patient A may have minor ones whil patient B may have extraordinary long blocks. We think that, in the ear of the listener, this difference in timing will be important. The downside of timing with this level of precision is that small measurement errors will occur (we are ony human), and this, in turn, will influence inter-rater unanimity.

Slide 32



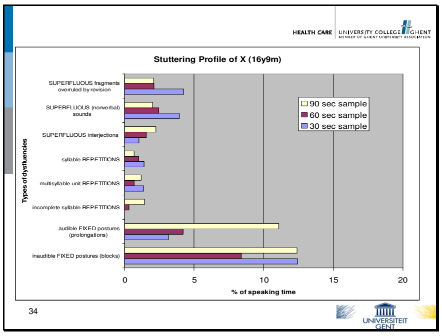
Anyhow, results come within seconds, but the annotation process was time consuming. Will we get the same result if we shorten the sample to win annotation time?
This is only a 30 seconds of speech, drawn from the start of the sample. This speaker predominantly has fixed postures and superfluous behaviors.

Slide 33



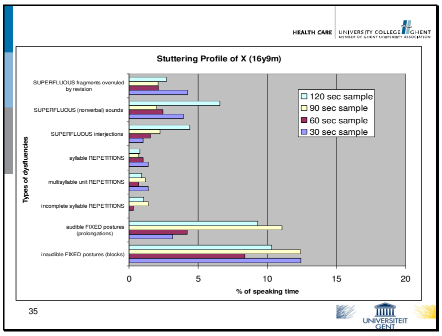
Nothing much changes by adding another 30 seconds to the analysis. This speaker predominantly has fixed postures and superfluous behaviors.

Slide 34



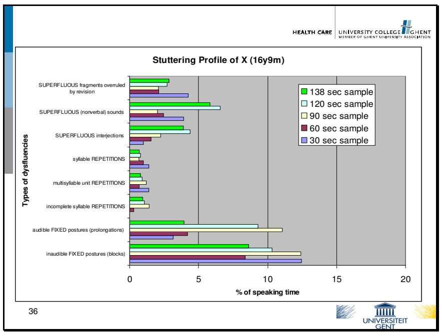
The same is true for the 90 sec sample. This speaker predominantly has fixed postures, only now a new type of fixed postures emerges.

Slide 35



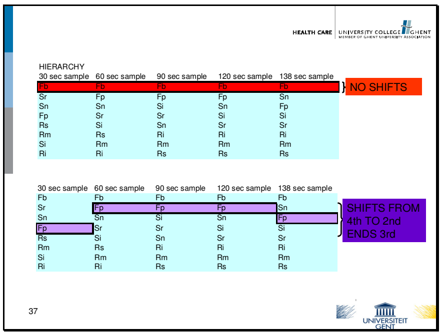
Two types of fixed postures and superfluous behavior.

Slide 36



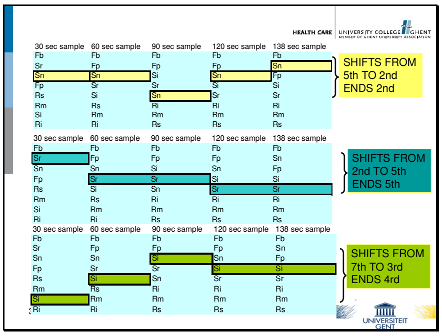
The same two types of fixed postures and superfluous behavior. This is about the same pattern that emerged from the shorter samples. Caveat, n=1! This cannot automatically be genaralized to samples from other speakers, more research is needed to answer that question. However, this may be a possibility to shorten annotation time. We could try to establish a rule, for instance: always analyse the worst 60 seconds and stop there. My hypothesis is that longer extracts will not reveal a different pattern in most speakers.

Slide 37



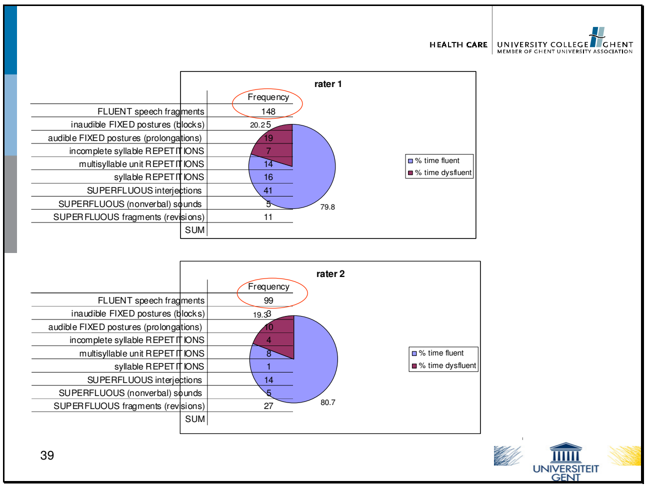
The fixed posture descriptors systematically end up in the top 2 of the hierarchy. They switch places but never jump more than 2 places within the hierachy. These speech related descriptors seem te be more stable than the other ones.

Slide 38



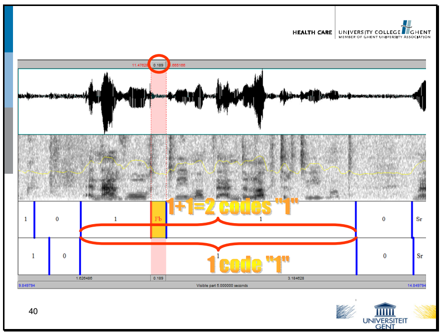
Except for the SUPERFLUOUS BEHAVIOR_nonverbal sounds, these descriptors do not end in the top 3. One could argue that this SUPERFLUOUS BEHAVIOR_nonverbal sounds descriptor is more speech related than the following ones. The other ones are more linguistic in nature, that is: to identify them one needs concepts from semantics and grammar. This is probably why they are influenced more by the length and the content of the sample. The SUPERFLUOUS BEHAVIOR_interjections descriptor jumps 4 places within the hierachy.

Slide 39



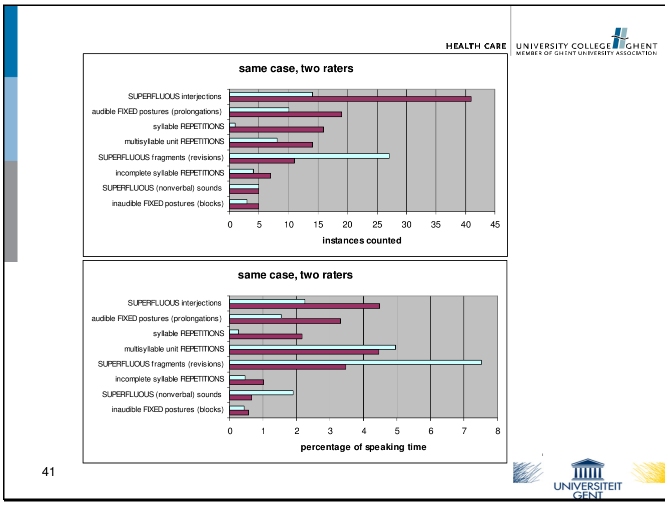
More research is needed to reveal the ideal sample duration. The same goes for inter-rater agreement. We did a small scale experiment with another sample (222 seconds, i.e. almost 4 minutes). It was processed by two raters. Both agreed very well on the degree of fluency vs dysfluency (about 80% of speaking time was considered fluent by both). The inter-rater differences were about the distribution of dysfluency types. The striking point here is that when we switch to counting instead of tallying, one of the raters finds a lot more instances of fluency (148 vs 99). This is in contrast to he unanimity in timing the amount of fluency. There is a simple explanation for this paradox.

Slide 40



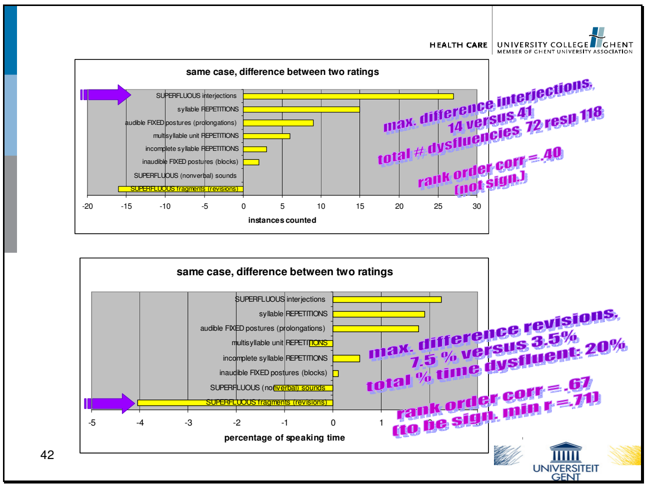
The TextGrids from both raters have been superimposed on the screen. Here you see the reason for the difference in tallied fluent instances. The first rater (TextGrid above) discerns a very short fixed posture of about 2 tenths of a second. The second rater (below) did not notice it. When you replay it loud enough you can hear it is a very tense swallow in the middle of an utterance. Video images may show it better. It is very short (0.2 s)(remember the total speaking time was 222 s) and therefore it does not really influence timing results. It does however influence countig results, since the original count of 1 fluent instance is now doubled. Obviously, the first rater does the job more thouroughly and there are more examples of this type of inter-rater disagreement. All these splitted fluent stretches result in a total of 148 for on of the raters,in contrast to a total of 99 fluent stretches for the other rater. The important point is that this disagreement is an artefact of tallying, not of timing.

Slide 41



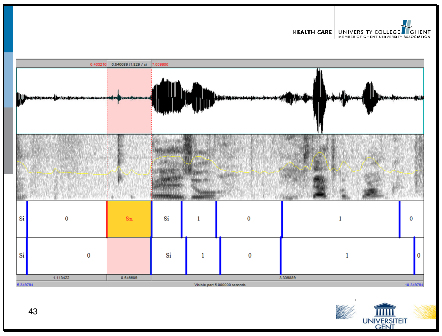
The inter-rater differences were not in the amount of fluency (80%) but rather in the distribution of dysfluency types within the remaining 20% of speaking time. Disagreement is most noticeable for superfluous behaviors (interjections and revisions) and for single syllable repetitions.

Slide 42



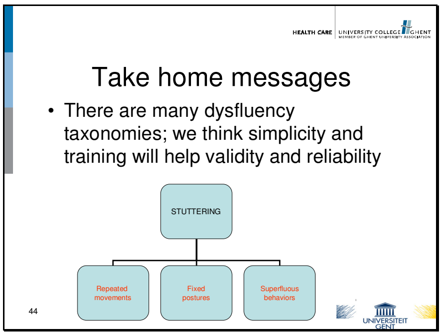
It is interesting to see that the level of disagreement between the two raters is different for tallied and timed results. When counting, interjections yields the biggest contrast. When timing, revisions yield the maximum contrast. These differences could be lessened by better instructions and identical/standard training before doing annotations. So maybe we need some sort of “indoctrination” here, a mandatory training with typical audiovisual examples of each category in the taxonomy. The same problem and a comparable solution can be found in the realm of voice disorders (how to rate GRBAS). Note that the rank order correlations between both raters’ annatoations are not significant. However, the correlation for timing measures is better and near significance thresholds (significance threshold here with n=8 is a r of at least .72). Again: this could be raised and become significant after annotation training.

Slide 43



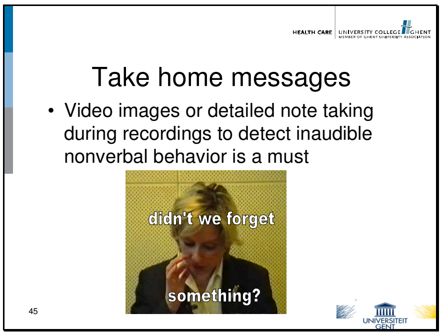
Another typical instance of disagreement between both raters: the speaker smacks hips lips before an utterance. For one of the raters this is superfluous nonverbal behavior. For the other this is the continuation of an otherwise irrelevant stretch of the recording. This is something that could have been avoided bij inspecting the oscillogram and the intensity curve and by replaying that part loud enough. Again: one of the raters was more thourough and this results in more annotations and a better resolution of the TextGrid.

Slide 44

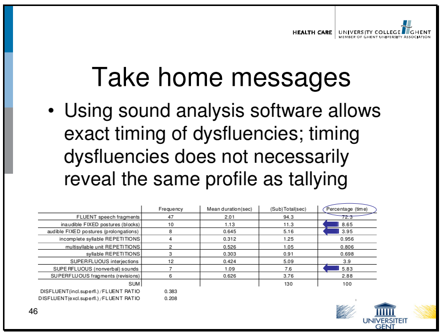


Stuttering descriptors that relate to motor speech behavior can be defined more unequivocally and we hypothesize that they tend to be “immune” to sample length. Some sort of standard training with typical audiovisual examples before doing annotations could very well improve the unanimity of annotators, i.e. inter-rater reliability.

Slide 45



Slide 46

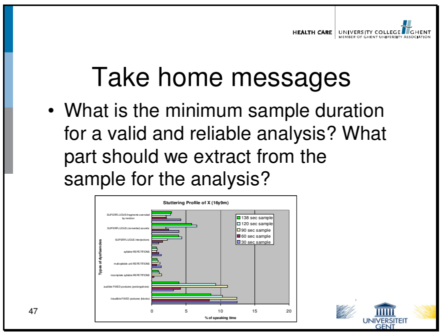


	Frequency	Mean duration(sec)	(Sub)Total(sec)	Percentage (time)
FLUENT speech fragments	47	2.01	94.3	72.3
inaudible FIXED postures (blocks)	10	1.13	11.3	8.65
audible FIXED postures (prolongations)	8	0.645	5.16	3.95
incomplete syllable REPETITIONS	4	0.312	1.25	0.956
multisyllable unit REPETITIONS	2	0.526	1.05	0.806
syllable REPETITIONS	3	0.303	0.91	0.698
SUPERFLUOUS interjections	12	0.424	5.09	3.9
SUPERFLUOUS (nonverbal) sounds	7	1.09	7.6	5.83
SUPERFLUOUS fragments (revisions)	6	0.626	3.76	2.88
SUM			130	100

DISFLUENT(incl.superfl.)/FLUENT RATIO 0.383
DISFLUENT(excl.superfl.)/FLUENT RATIO 0.208

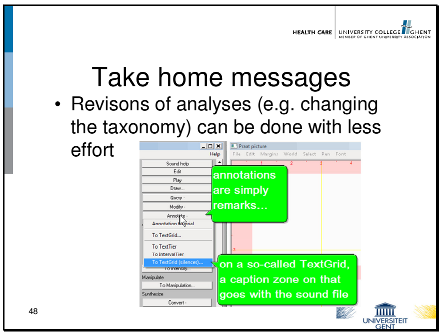
Tallying and timing does not result entirely in the same outcome. For prolongations and blocks this difference is obvious: two patients may have the same frequency but the may very well differ drastically in duration of blocks and prolongations. Also, it may very well be that timing measures are more sensitive for subliminal changes in behavior. For example, as a result of therapy a stutterer may not be able limit the number of fixed postures or superfluous behavior, but he may be able to shorten these episodes, even without the therapist being aware of it. Finally, the border between normal fluency and stuttering could be reformulated in seconds, which may allow more granularity.

Slide 47



Perhaps there is a possibility to shorten annotation time. We could try to establish a rule, for instance: always analyse the worst 60 seconds and stop there. My hypothesis is that longer extracts will not reveal a different pattern in most speakers.

Slide 48



Remember the long pauses for inhaling? At first you may code them as irrelevant fragments (“0”), but if you decide to treat them as superfluous nonverbal behavior, one can simple change the text grid mnemonics and recalculate. Don’t forget to save the text grid on your harddisk!