Processus de validation du test SELF (Système d'Évaluation en Langues à visée Formative) japonais - premiers pilotages des items

Sylvain Coulange
Université Grenoble Alpes, Innovalangues

Résumé :

Cet article présente les deux premières sessions de pilotage du test du système d'évaluation en langues SELF pour le japonais. Il y sera détaillé l'organisation des passations du test, les participants, le processus de traitement des données et enfin les résultats obtenus. Ces pilotages se situent eux-mêmes au cœur d'un long processus de validation du test de positionnement, détaillé en introduction.

Mots-clés : positionnement, évaluation, test, SELF, japonais, CECRL

Introduction

Un grand nombre d'étapes de validations sont nécessaires pour garantir la fiabilité du test SELF. Le processus de validation est identique pour toutes les langues du projet. Une fois les items rédigés par des enseignants concepteurs, une première réunion de validation a lieu en équipe. Les items sont alors modifiés ou supprimés si besoin. Après avoir atteint un certain nombre d'items, un test de pilotage est organisé, il permet de tester les items auprès d'une population d'apprenants pré-positionnés. À l'issu du pilotage, les items sont à nouveau triés en fonction des résultats obtenus ; si une modification doit avoir lieu, l'item est alors testé à nouveau lors d'un pilotage ultérieur. Une fois un nombre suffisant d'items validés réunis, une troisième grande étape de validation a lieu : le pré-test. Celui-ci doit tester à nouveau les items sur une population plus importante que le pilotage, et présente les items des tous les niveaux réunis aux apprenants. Une ultime sélection des items est réalisée, et il est alors possible de définir précisément les frontières de niveau, grâce à l'aide d'enseignants experts non-rédacteurs de différentes universités (s*tandard setting*).

Cet article se donne pour objectif de présenter les deux premières sessions de pilotage du test SELF japonais, ainsi que les résultats obtenus.

I. Les Pilotages

De septembre 2015 à juin 2016, deux sessions de pilotage ont été organisées. Ces pilotages ont pour objectif de valider une première fois les items conçus pour le test SELF Japonais. Ils se composent d'une succession d'items de même niveau, présentés à une population pré-positionnée. Le niveau des participants est donc connu, l'intérêt du pilotage est alors de vérifier si les items sont adaptés au niveau cible. L'université de Paris Diderot (Paris 7, P7) et l'Institut national des Langues et Civilisations Orientales (INALCO) et l'Institut national des Sciences Appliquées de Lyon (INSA) ont collaboré avec l'université Grenoble Alpes (UGA) pour recueillir un total de 196 passations de tests pilotes du SELF japonais de niveau A1 et A2. Un premier pilotage a eu lieu en janvier, avec 106 participants provenant de trois institutions, la moitié ayant passé un test A1, l'autre moitié un test A2. Un deuxième pilotage a été organisé en avril, réunissant 90 passations, dont 40 de niveau

A1 et 50 de niveau A2, dans les quatre institutions. L'illustration 1 présente la répartition des participants par niveau de test et par institution.

Pilotage 1

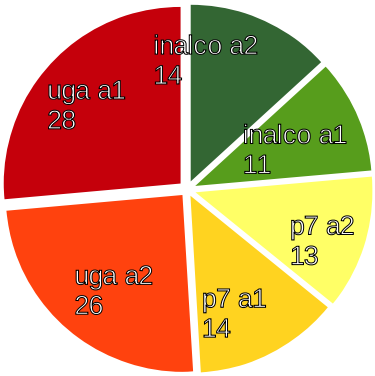
	Niveau du test	Nombre de participants
Grenoble Alpes	A1	28
	A2	26
Paris 7	A1	14
	A2	13
Inalco	A1	11
	A2	14

Total A1	*53*
Total A2	*53*
Total général	***106***

Pilotage 2

	Niveau du test	Nombre de participants
Grenoble Alpes	A1	18
	A2	24
Paris 7	A1	1
	A2	8
Inalco	A1	5
	A2	5
INSA	A1	16
	A2	13

Total A1	*40*
Total A2	*50*
Total général	***90***



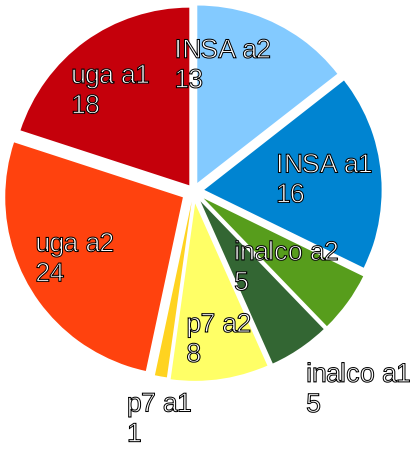


Illustration 1: Répartition des participants par test et par université

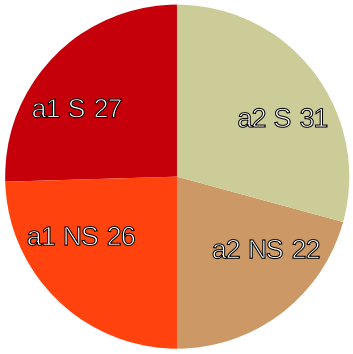
Parmi les 196 passations, 108 ont été réalisées par des apprenants non-spécialistes (NS) de japonais, et 88 par des apprenants spécialistes (S). La différence de proportion des participants dans les deux pilotages s'explique en partie par le fait qu'après 8 mois de cours intensifs, la plupart des spécialistes ont atteint le niveau A1, voire l'ont dépassé, et ils ne représentent donc plus une population en cours d'acquisition A1, comprenant des apprenants début-A1, mi-A1 et fin-A1. Tandis que la vitesse de progression d'apprentissage chez les non-spécialistes est beaucoup moins importante, et beaucoup d'étudiants sont encore en cours d'acquisition A1. Un nombre plus important de non-spécialistes ont donc été conviés à passer le pilotage 2 A1 pour palier à ce décalage et permettre une validation plus uniforme des items du test.

Pilotage 1

	Non-spécialistes	Spécialistes
A1	26	27
A2	22	31
Total	**48**	**58**

Pilotage 2

	Non-spécialistes	Spécialistes
A1	31	9
A2	29	21
Total	**60**	**30**



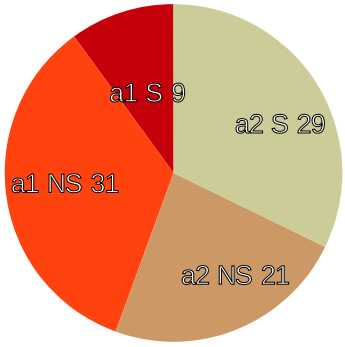
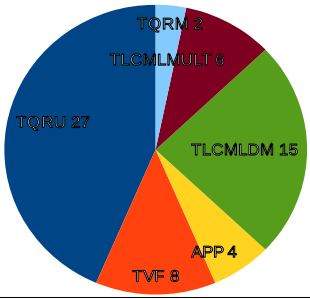
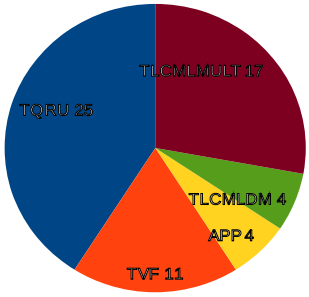


Illustration 2 : Répartitions des participants spécialistes/non-spécialistes

Afin d'effectuer une première validation des items du SELF Japonais, nous proposons des tests sans étape. Les participants sont donc confrontés à l'ensemble des items constituant le test d'un même niveau, dans un ordre prédéfini et identique pour toute la population de testeurs. Les items sont contenus dans des tâches, triées selon la compétence testée. Chaque test débute par les tâches de compréhension orale (CO), puis écrite (CE) et se termine par les tâches d'expression écrite courte (EEC).

Les items peuvent avoir différents types : il peut s'agir d'une question à choix unique (TQRU), majoritaire, d'une question à choix multiple (TQRM), de listes multiples (TLCMLMULT) ou uniques (TLCMLDM), d'appariement (APP) ou encore de vrai-faux (TVF). L'illustration 3 présente la constitution détaillée des tests.

	Pilotage1_A1			Pilotage1_A2		
Tâches	45			35		
	CO : 18	CE : 13	EEC : 14	CO : 14	CE : 10	EEC : 10
Items	62			61		
	CO : 18	CE : 20	EEC : 24	CO : 19	CE : 20	EEC : 22
Types d'items						

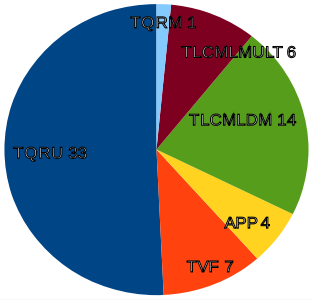
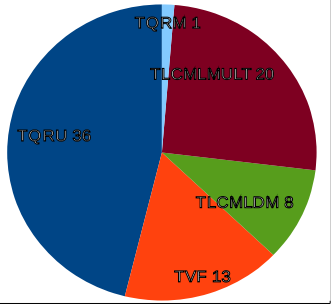
	Pilotage2_A1			Pilotage2_A2		
Tâches	48			45		
	CO : 19	CE : 14	EEC : 15	CO : 15	CE : 18	EEC : 12
Items	65			78		
	CO : 19	CE : 21	EEC : 25	CO : 20	CE : 30	EEC : 28
Types d'items						

Illustration 3 : Constitution des tests

Cinq passations du premier pilotage ont été réalisées en *think aloud*, pour les niveaux A1 et A2 (*cf.* Higashi, dans les mêmes actes). Ces passations ont permis d'avoir un aperçu de la réflexion des participants en temps réel, pendant la passation du test. Une grande quantité de données ont pu être rassemblées ; et nous ont aidé à améliorer la rédaction des items, leur présentation graphique, ainsi qu'à comprendre pourquoi certains items étaient plus problématiques que d'autres. Les résultats aux tests ont été ajoutés à ceux des passations classiques.

II. Analyse quantitative des données

Après chaque pilotage, nous obtenons la liste des réponses choisies par chaque participant pour chaque item. Nous y associons la liste des items pour chaque test ainsi que la ou les clés correspondantes (réponses correctes). Le logiciel TiaPlus permet de mettre en relation ces données brutes afin de déterminer un certain nombre d'informations. Pour la validation post-pilotage, nous nous concentrons sur l'analyse des données suivantes : l'indice de difficulté de l'item (P)[1], la corrélation entre la réussite à l'item et la réussite au test (Rir)[2], le pourcentage de choix de chaque option de réponse par item, ainsi que le coefficient alpha de *Cronbach* pour mesurer la cohérence interne des items dans le test. Nous observons aussi le pourcentage de choix de chaque option de réponse en fonction du résultat global au test, pour déterminer quelle option est choisie, et par qui.

La population de testeurs est automatiquement divisée en quatre groupes, en fonction du résultat global au test (P global). Le groupe 1 est celui qui a obtenu les moins bons résultats globaux au test, le groupe 4 est celui ayant obtenu les meilleurs résultats. En affichant pour chaque item, les options de réponse choisies par les participants, on peut alors obtenir des courbes de réponses qui représentent par extrapolation l'évolution du choix en fonction du niveau de japonais du participant. Dans l'illustration ci-contre, pour un item donné, on constate que 0 % des testeurs appartenant au sous-groupe de moins bon niveau (ayant le moins bon score global) ont choisi la bonne réponse

1 Celui-ci n'est autre que le pourcentage de réponses correctes sur l'ensemble des réponses.
2 Le Rir exclue le résultat de l'item en question lors de la corrélation avec le résultat global du test, le Rit quant à lui l'inclue au résultat global.

(indiquée par * dans la légende), tandis que 100 % des testeurs du sous-groupe le plus avancé l'ont choisi. Quant aux distracteurs (réponses A et C), leur sélection décroît à mesure que le niveau du sous-groupe augmente. On a donc ici un item particulièrement discriminant, très adapté à la population testée et dont les distracteurs sont efficaces. À l'inverse, si la courbe de choix de la bonne réponse décroît au fur et à mesure que le niveau de la population augmente, l'item est inutilisable car mal discriminant. En outre, si la courbe de choix d'un distracteur est horizontale ou ascendante, cela signifie que son choix ne décroît pas en fonction de la progession de l'apprentissage ; on doit donc réviser le distracteur, ou le supprimer[3].

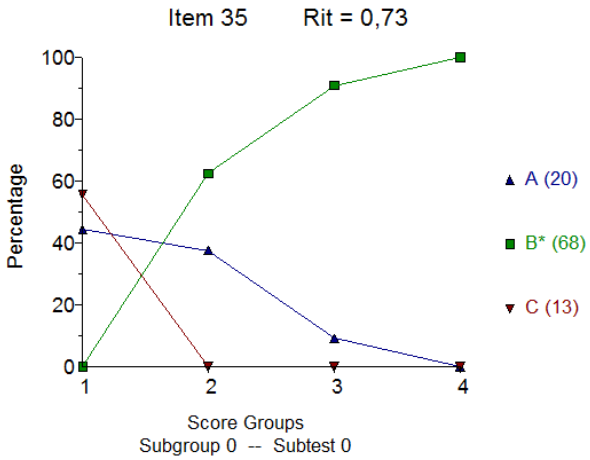


Illustration 1: Choix de réponse en fonction du sous-groupe de testeurs

La plate-forme SELF que nous utilisons pour les passations permet d'exporter d'autres données, comme le temps passé par tâche et par participant, la difficulté perçue par le participant pour chaque tâche (sur une échelle de 5) ou des informations relatives au profil du participant (université, filière, année, nom etc.)[4]. C'est le croisement de l'ensemble de ces données, en plus du contenu pédagogique des items, qui nous permet de valider, réviser ou supprimer chaque item du test. La validation se passe en équipe (5 personnes), et passe en revue chaque item afin de ne garder que ceux qui sont cohérents, fidèles au test et dont le taux de discrimination entre les bons et moins bons apprenants est élevé. Puisque c'est là tout l'intérêt du test : mesurer la différence de connaissance du japonais entre les participants et les placer le plus précisément possible sur une échelle de progression basée sur les descripteurs du CECRL.

Comme le montre l'illustration 2, les 62 items du niveaux A1 du premier pilotage ont été testés en une moyenne de 52 minutes, sans compter les cinq passations en *think-aloud*, qui ont duré environ 1h30 chacune. On constate que le score global moyen obtenu s'élève à 82 % de réussite. Les scores moyens des trois compétences sont les suivants : 88 % en compréhension orale, 81% en compréhension écrite et 80 % en expression écrite courte. On constate que les pourcentages de réussite par item sont élevés : 23 items sont réussis par plus de 90 % de la population, dont 2 par 100 % des participants. La version finale du test SELF présentera aux participants des items de niveaux supérieurs en plus des items A1, ce qui entraînera une légère baisse du score global. Le pouvoir de discrimination moyen des items (Rir moyen) est de 0,42, soit un pouvoir très discriminant[5].

Au deuxième pilotage, 65 items ont été utilisés pour le niveau A1. La passation a été plus longue d'une moyenne de 13 minutes par rapport au premier pilotage. Les résultats sont globalement inférieurs : 72 % pour l'ensemble du test, 84 % pour la compréhension orale, 69 % pour la compréhension écrite et 64 % pour l'expression écrite courte. Le pouvoir de discrimination moyen

3 Toute modification de l'item nécessite son repilotage.

4 Les participants sont amenés à évaluer la difficulté de l'exercice sur une échelle de 5, après chaque tâche. Cette évaluation n'a lieu que pendant les sessions de pilotage.

5 D'après Jean-Marc Braibant, de l'université catholique de Louvain, dans sa note sur les examens QCM (https://www.uclouvain.be/cps/ucl/doc/adef/documents/EVA_QCM_version3.pdf)

est également inférieur : 0,34. On a donc un test globalement plus dur que le premier, nécessitant plus de temps de passation, et discriminant moins bien les participants en fonction de leur niveau de japonais, si l'on admet que la population est représentative des apprenants de ce niveau et que leur score global au test correspond à leur niveau général de japonais.

	Pilotage 1 – test A1	
Temps moyen de passation	51,9 minutes	
Score moyen	82%	CO : 88 % CE : 81 % EEC : 80 %
Pouvoir de discrimination moyen des items (Rit)	0,42	

	Pilotage 1 – test A2	
Temps moyen de passation	56,5 minutes	
Score moyen	80%	CO : 93% CE : 82% EEC : 67%
Pouvoir de discrimination moyen des items (Rit)	0,39	

	Pilotage 2 – test A1	
Temps moyen de passation	64,8 minutes	
Score moyen (P)	72%	CO : 84% CE : 69% EEC : 64%
Pouvoir de discrimination moyen des items (Rit)	0,34	

	Pilotage 2 – test A2	
Temps moyen de passation	85 minutes	
Score moyen (P)	67%	CO : 85% CE : 68% EEC : 54%
Pouvoir de discrimination moyen des items (Rit)	0,33	

Illustration 2: Résultats globaux obtenus aux deux pilotages

En ce qui concerne le niveau A2, le test du premier pilotage a été effectué en une moyenne de 57 minutes. Le score global est de 80 %, 93 % en CO, 82 % en CE et 67 % en EEC. Les résultats sont assez proches de ceux obtenus au niveau A1 lors du même pilotage, mise à part l'expression écrite. Le pouvoir de discrimination moyen des items est de 0,39.

Pour le pilotage 2 au niveau A2, on observe là encore des résultats globalement inférieurs et un temps moyen de passation particulièrement long. Une moyenne de 85 minutes ont été nécessaires pour terminer le test ; le résultats global est de 67 %, avec 85 % pour la CO, 68 % pour la CE et 54 % pour l'EEC. Le pouvoir de discrimination est quant à lui de 0,33. On observe donc la même tendance que pour le niveau A1, le premier pilotage est globalement plus réussi et plus discriminant ; tandis que le second pilotage est plus long, plus fastidieux et moins discriminant. Nous avons vu dans la partie précédente qu'il y a effectivement plus d'items dans les deux niveaux du second pilotage, ce qui peut expliquer le temps de passation. Le niveau de la population de testeurs était plus homogène qu'au premier pilotage (l'écart entre les niveaux des participants est moins important), ce qui peut être à l'origine d'une discrimination moins précise.

Les résultats obtenus en compréhension orale, autant pour le test de niveau A1 que celui de niveau A2, sont assez proches de ceux obtenus au premier pilotage. Entre les scores moyens obtenus aux deux pilotage, on a seulement une différence de 2 % en A1 et 6 % en A2, avec toujours un résultat inférieur au deuxième pilotage. Cependant la différence des scores entre les pilotages pour les autres compétences est bien plus importante (12 % et 16 % en A1 ; 13 % et 13 % en A2,pour la CE et l'EEC respectivement). Cette différence peut être expliquée par le fait que les items de CE et EEC du pilotage 2 ont été rédigés après la validation des items du premier pilotage ; tandis que les items de CO ont tous été réalisés avant le premier pilotage. Autrement dit, nous avons conçu les

items de CE et EEC dans l'objectif de créer des items plus complexes, étant donné que beaucoup d'items simples avaient été validés au premier pilotage. Pour la CO cependant, tous les items ont été créés en même temps, sans viser un sous-niveau spécifiquement, c'est donc la raison pour laquelle la différence de score en CO entre les deux pilotage est comparable. Quoiqu'il en soit, le facteur population de testeurs est toujours à garder à l'esprit, il est lui aussi évidemment à l'origine de variations importantes des résultats. Après de nombreuses remarques d'étudiants et d'enseignants au premier pilotage, nous avons aussi décidé d'augmenter légèrement la vitesse de diction des items de compréhension orale (*cf.* Shirota, dans les mêmes actes).

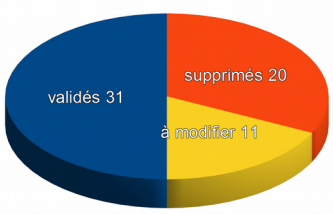
Après avoir procédé à une analyse détaillée des items, en croisant l'ensemble des données et métadonnées obtenues, nous avons réparti les items en trois catégories : "À garder" / "À modifier" / "À jeter". Seuls les items de la première catégorie seront conservés tels quels, et testés à nouveau sur une plus grande population lors du prétest, dernière étape de validation avant la livraison du test.

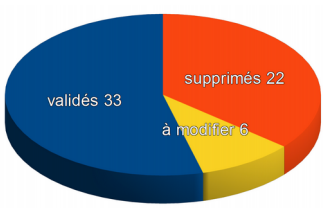
	Pilotage 1 – test A1			
	CO	CE	EEC	**Total**
À garder	6	7	18	**31**
À modifier	3	6	2	**11**
À jeter	9	7	4	**20**
			total :	**62**

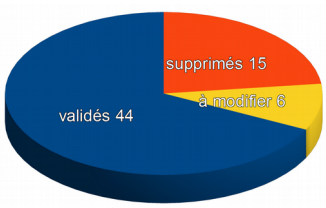
	Pilotage 1 – test A2			
	CO	CE	EEC	**Total**
À garder	7	8	18	**33**
À modifier	0	6	0	**6**
À jeter	12	6	4	**22**
			total :	**61**

	Pilotage 2 – test A1			
	CO	CE	EEC	**Total**
À garder	10	16	18	**44**
À modifier	5	1	0	**6**
À jeter	4	4	7	**15**
			total :	**65**

	Pilotage 2 – test A2			
	CO	CE	EEC	**Total**
À garder	7	27	23	**57**
À modifier	5	1	0	**6**
À jeter	8	2	5	**15**
			total :	**78**







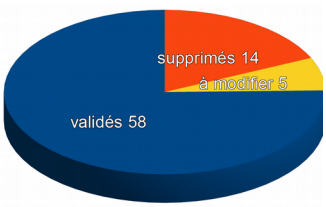


Illustration 3: Résultats obtenus après validation post-pilotage

Comme le présente l'illustration 3, le nombre d'items validés est allé croissant en fonction des pilotages. Un total de 75 items de niveau A1 ont été gardés tels quels, et 90 items de niveau A2. Les items de compréhension orale ont été peu efficace (score global élevé, pouvoir discriminant bas), beaucoup ont donc été supprimés. À l'inverse, la plupart des items d'expression écrite ont été très discriminant, un grand nombre d'entre eux ont donc pu être conservés. En fonction des résultats pour chaque item, ainsi que du niveau des participants, nous avons réparti les items validés sur trois sous-niveaux pour A1 et A2. L'illustration 4 présente la proportion d'items validés en fonction du sous-niveau attribué. Ces sous-niveaux sont susceptibles d'évoluer en fonction des résultats obtenus à plus grande échelle, lors du prétest. Grâce à ces données, nous pouvons prévoir quels items doivent être rédigés par la suite, pour compléter la banque de données du SELF Japonais.

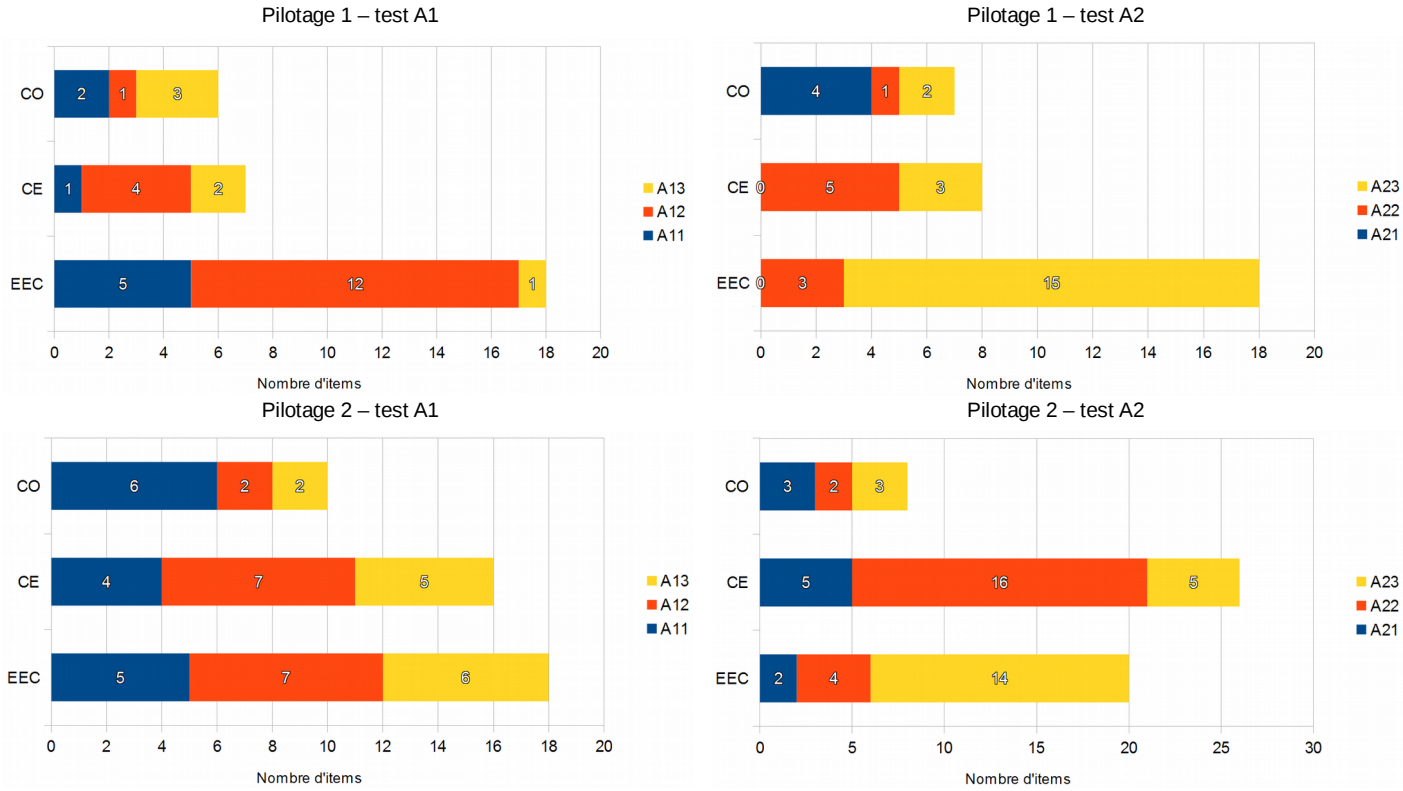


Illustration 4: Détails des sous-niveaux des items validés

Conclusion :

C'est un long procédé de validation qui est nécessaire pour assurer la fiabilité d'un test de langue. Nous avons eu bien-sûr besoin d'un grand nombre d'apprenants cobayes, mais aussi de l'aide de plusieurs enseignants experts et extérieurs, pour nous apporter un regard objectif sur le test, et les items qui le composent. Les tests pilotes tels qu'ils ont été réalisés ici, nous ont permis de savoir comment a été effectué chaque item du test, et par quels étudiants. On distingue tout de suite une échelle de progression le long de laquelle se situent les apprenants, en fonction de leurs résultats. Les items présentant des données incohérentes avec le groupe d'apprenants sont identifiés, puis corrigés ou supprimés ; pour ne laisser finalement que des items à haut taux de discrimination. Il faut croiser un grand nombre de données pour juger au mieux de l'efficacité d'un item, tout en étant vigilent aux possibles facteurs externes pouvant biaiser certains résultats (logiciel, conditions de passation etc.).

Une fois qu'une banque d'items vérifiés, discriminants et en nombre suffisant pour évaluer chaque niveau du CERCL de A1 à B1, le prétest pourra être organisé. Il permettra alors de définir le plus précisément possible le niveau de chaque item, et ainsi préciser le niveau des apprenants qui les réussiront.